Paulo Sérgio Fonte Ferreira

A METRIC APPROACH TO SEX ESTIMATION USING THE HUMERUS, THE ULNA AND THE RADIUS

Dissertação no âmbito do mestrado em Evolução e Biologia Humanas, orientada pela Professora Doutora Cláudia Umbelino e pelo Professor Doutor Francisco Curate e apresentada ao Departamento de Ciências da Vida da Faculdade de Ciências e Tecnologia da Universidade de Coimbra.

outubro de 2021

Paulo Sérgio Fonte Ferreira

A metric approach to sex estimation using the humerus, the ulna and the radius

Dissertação no âmbito do mestrado em Evolução e Biologia Humanas, orientada pela Professora Doutora Cláudia Isabel Soares Umbelino e pelo Professor Doutor José Francisco Taborda Curate e apresentada ao Departamento de Ciências da Vida da Faculdade de Ciências e Tecnologia da Universidade de Coimbra.

Outubro de 2021

Acknowledgments

Those who know me know that I'm not one for many words.

First of all, I thank my mother, my sister and my grandmother for their unwavering support and encouragement throughout the years. Without their sacrifices I wouldn't be where I am today.

To my advisors, Professor PhD Francisco Curate and Professor PhD Cláudia Umbelino, for the encouragement and the complete autonomy afforded. I also want to thank the help and availability of Professor PhD Ana Luísa Santos.

A note of thanks to the Department of Life Sciences (DCV/UC) and Professor PhD Sofia Wasterlain for the availability and access to the Coimbra Identified Skeletal Collection (CISC) and to the Laboratory of Forensic Anthropology (LFA/CFE) and Professor PhD Maria Teresa Ferreira for the opportunity to collect data from the 21st Century Identified Skeletal Collection (CEI/XXI).

To my friends Pedro, Sara and Ivan and a few others goes a big thank you for the companionship, the conversations and discussions, and motivation. To my international family, Isabel, Alex, Ricciardo, Vito and Giulia, I feel very grateful for the comradery and the great memories we share.

Finally, a general note of appreciation to all my friends and professores that, one way or another, had an impact on me during these formative years.

Abstract

Sex estimation is one of the most important steps in the process of building the biological profile of an unidentified individual. In decomposing or skeletonized individuals, bone analysis might provide the only way to access biological sex. When more dimorphic bones are not available or in case of fragmentary or incomplete remains, a combination of measurements from the humerus, ulna and radius can be used to develop methods for sex estimation.

Measurements of the left the humerus, ulna and radius were taken from a sample of 280 adult individuals (140 males and 140 females) from the Coimbra Identified Skeletal Collection (late 19th - early 20th centuries) in order to generate univariate and multivariate models for sex estimation. All models were evaluated with a 10-fold cross-validation method and tested on an independent sample composed of 50 adult individuals (22 males and 28 females) from the 21st Century Identified Skeletal Collection (late 20th - early 21st centuries), also housed in Coimbra.

Univariate models show accuracies ranging from 75.4% to 94.4% (cross-validation), and from 60.5 to 82.4% (test sample), while accuracy in multivariate models varies from 89.4% to 96.7% (cross-validation), and 74.3% to 86.4% (test sample). Overall the best results, when tested on the holdout sample, for uni and multivariate models resulted from the use of measurements from the radius, with accuracies ranging from 70.6% to 82.4% and 81.8% to 86.4%, respectively.

Results suggest that measurements of the long bones of the upper limbs are useful to develop methods for sex estimation of anonymous skeletal remains. However, attention is required when employing such methods on more recently deceased individuals.

Keywords: forensic anthropology; identified skeletal collection; sex diagnosis; biological profile; osteometry

Resumo

A estimativa do sexo é um dos passos mais importantes na criação do perfil biológico de indivíduos não identificados. Em indivíduos em decomposição ou esqueletizados, a análise óssea é uma das únicas formas de aceder ao sexo biológico. Na ausência de ossos mais dimórficos, ou no caso de restos fragmentários ou incompletos, a combinação de medidas do úmero, da ulna e do rádio podem ser usadas para desenvolver métodos de diagnose sexual.

Medidas do úmero, ulna e rádio esquerdos foram recolhidas de uma amostra de 280 indivíduos adultos (140 masculinos e 140 femininos) da Coleção de Esqueletos Identificados de Coimbra (fim séc. XIX - início séc. XX) de forma a desenvolver modelos univariados e multivariados para a estimativa do sexo. Todos os modelos foram avaliados com um método de validação cruzada e testados numa amostra independente composta por 50 indivíduos adultos (22 masculinos e 28 femininos) da Colecção de Esqueletos Identificados do séc. XXI (fim do séc. XX - início do séc. XXI), também alojados em Coimbra.

Os modelos univariados revelaram uma precisão que varia entre 75,4% a 94,4% (validação cruzada), e de 60,5 a 82,4% (amostra de teste), enquanto a precisão nos modelos multivariados varia entre 89,4% a 96,7% (validação cruzada), e 74,3% a 86,4% (amostra de teste). Na generalidade, quando usados na amostra de teste, os métodos uni e multivariados que usam medidas do rádio mostram precisões de 70.6% a 82.4% e 81.8% a 86.4%, respectivamente.

Os resultados sugerem que as medidas dos ossos longos do membro superior são úteis para desenvolver métodos de estimação sexual de restos esqueléticos não identificados. No entanto, é necessária atenção ao empregar estes métodos em indivíduos falecidos mais recentemente.

Palavras-chave: antropologia forense; coleção de esqueletos identificados; diagnose sexual; perfil biológico; osteometria

Table of Contents

Index of Tables

Index of Figures

List of Abbreviations

CISC	Coimbra Identified Skeletal Collection
CEI/XXI	21st Century Identified Skeletal Collection
TEM	Technical Error of Measurement
LR	Logistic Regression
SVM	Support Vector Machines
SDI	Sexual dimorphism index

Measurements

HML	Maximum length of the humerus
HVDH	Vertical diameter of the head of the humerus
HHDH	Horizontal diameter of the head of the humerus
HMC	Minimum circumference of the humerus
HEB	Epicondylar breadth of the humerus
HCB	Condylar breadth of the humerus
UML	Maximum length of the ulna
UPL	Physiological length of the ulna
UMC	Minimum circumference of the ulna
UOB	Olecranon breadth of the ulna
UBSN	Breadth of the semilunar notch of the ulna
UDAW	Distal articular width of the ulna
RML	Maximum length of the radius
RMCDD	Minimum circumference of the radius
RHC	Head circumference of the radius
RNC	Neck circumference of the radius
RDEW	Distal epiphysis width of the radius

1. Introduction

The present dissertation aims to develop and test metric sex estimation methods based on the humerus, ulna and radius from a Portuguese identified skeletal collection. Thus, this work pretends to build upon the existing methods for sex estimation and provide new insights.

1.1. Objectives

- Obtain a set of measurements from the humerus, ulna and radius;
- Assess the most dimorphic regions of each bone;
- Develop metric methods for sex estimation for each of the bones individually and grouped
- Observe, if present, the effects of secular trends in the accuracy of the methods

1.2. Importance of Sex Estimation

Humans display a pattern of morphological differentiation between males and females, termed sexual dimorphism, that makes biological sex one of the primary pieces of information that researchers seek to obtain from skeletonized remains. Sex estimation is an important element in the establishment of the biological profile along with age-at-death (Márquez-Grant, 2015), ancestry (Algee-Hewitt et al., 2020; Christensen et al., 2019; Navega, Coelho, et al., 2015) and stature (Zeman & Beňuš, 2020).

Paleoanthropologists, bioarchaeologists and forensic anthropologists alike are interested in estimating the sex of any recovered human remains (Bethard & VanSickle, 2020).

In the field of paleoanthropology it allows the study of questions related to sexual dimorphism, the social systems and morphological adaptations of our hominin ancestors from an evolutionary lens. Relevant examples of the morphological changes that interest paleoanthropologists are the evolution of the pelvis (Rosenberg & DeSilva, 2017), as it relates to parturition or bipedalism, and of the skull, with its relation to the increased brain-size observed in our evolutionary lineage (Bruner et al., 2003; Gunz et al., 2010). Limitations arise, however, seeing as methods developed on modern humans are not guaranteed to be applicable to hominin fossil remains. As for bioarchaeologists, they couple sex estimation with contextual information retrieved from excavation sites (Zuckerman & Crandall, 2019), allowing for a better understanding of the paleodemography, mortality and sex-specific activities of past populations. In this case an incorrect classification can fundamentally alter the interpretation and context of an archaeological site (Messer & Getz, 2020).

For forensic anthropologists the estimation of biological sex from anthropological analysis of decomposed and skeletonized human remains represents a key element in the process leading to a positive identification (Pinheiro & Cunha, 2006). The accurate estimation of sex narrows the focus of the investigation, helping law enforcement, medical examiners and medicolegal death investigators narrow the list of potential unidentified persons (Bethard & VanSickle, 2020).

One must always bear in mind that these methods tend to attribute a binarism to human sex and do not always represent the spectrum of variability that it contains.

1.3. Sexual Dimorphism and Population Variation

The differences observed between male and female individuals are mainly associated with primary and secondary sexual characteristics (Black et al., 2009). The primary sexual characteristics are related to reproduction, with the differences in the pelvic morphology as an indicator; while the secondary sexual characteristics are unrelated to reproduction, for example differences in body height and weight that are influenced by genetic and hormonal differences (Nikitovic, 2018). At the end of adolescence males tend to exhibit greater stature and weight, however the actual values have great global variation (Ubelaker & DeGaglia, 2017).

Social phenomena such as the sexual division of labor (Ruff, 1987), resource acquisition, or parental investment can significantly affect the levels of sexual dimorphism for different populations. Physically demanding activities that are practiced asymmetrically between the sexes can exacerbate differences and increase sexual dimorphism (Berner et al., 2017), whereas if practiced equally would reduce it.

The level of sexual dimorphism exhibited by a population is the consequence of a collection of factors, such as the local ecological context (Wells, 2012), social behaviour, and genetics (Betti, 2017), with different levels of expression. This results in differences in dimorphism observed across different populations and contexts. The effects of secular trends are also important and can have great influence in the accuracy of sex estimation methods (Gonçalves, 2014; Langley & Jantz, 2020).

1.4. Methodological Approaches and Identified Skeletal Collections

Sex estimation of human skeletal remains involves metric approaches, visual or morphological observation and molecular methods. The availability of identified and documented human skeletal collections provided the necessary data for the development and study of these methods. The Hamman-Todd collection in Cleveland, Ohio and the Terry Collection in Washington, DC, were early collections that provided fertile ground for this type of research (Ubelaker & DeGaglia, 2020).

However, methods developed from these collections, using North American populations, were not representative of the global sexual dimorphism variation encountered, with uncertainty regarding their accuracy when applied to different populations. New and diverse documented collections were and are currently being established in different parts of the world with local populations (Petaros et al., 2021; Ubelaker, 2014).

In Portugal examples of this trend are notorious with the establishment of the Coimbra Identified Skeletal Collection (CISC) (Cunha & Wasterlain, 2007), which is the basis of this dissertation, made up of skeletal remains from individuals born between 1817 and 1924 and the 21st Century Identified Skeletal Collection also housed at the University of Coimbra (Ferreira et al., 2014, 2021). These collections provide information on individuals from a wide temporal range.

1.4.1. Methods for sex estimation: state-of-the-art

Traditional morphological approaches for sex estimation focus on the pelvis and the skull. As indicated above, because of the reproductive differences, the pelvis is highly dimorphic, and as such does not require a very fine observation to separate males from females. Similarly, the skull also possesses multiple morphological traits that are sexually dimorphic. Research has shown that the pelvis can provide accuracies up to 95% in sexing individuals, with the skull achieving accuracies between 70% to 80%. Methods for postcranial bones were developed for when the pelvis or skull are absent or fragmented (Spradley & Jantz, 2011). Methods for other postcranial bones have also been developed, such as the humerus (Ammer et al., 2019) and the clavicle (Rogers et al., 2000).

Metric sex determination methods (Spradley, 2016) exist for almost every bone of the human skeleton with a wide variety of statistical approaches, with the long bones proving very effective, with multivariate models of the postcranial bones showing accuracies of up to 94% (Spradley & Jantz, 2011). The availability of identified skeletal collections and the ease of data acquisition make this a useful method for anthropologists. One of the limitations of these methods is that they are usually population-specific, that is, they work best for individuals of the reference populations with accuracy declining for individuals from different contexts.

There are methods available for the pelvis (d'Oliveira Coelho & Curate, 2019), the cranium (Gillet et al., 2020), teeth (Peckmann et al., 2016), the femur (Curate et al., 2017; Curate et al., 2016; Cuzzullin et al., 2020), hand bones (DeSilva et al., 2014; El Morsi & Al Hawary, 2013; Falsetti, 1995), foot bones (Navega et al., 2015; Silva, 1995) and vertebrae (Gama et al., 2015). This is not an exhaustive list of the available methods, with many more existing for a variety of populations.

The long bones of the upper body, the humerus, ulna and radius, have a great record of providing accurate estimations of sex, with accuracies ranging from 87.4% to 97.5% (Albanese, 2013). Wasterlain (2000) using the vertical diameter of the head and epicondylar breadth of the humerus shows the potential for high accuracy sex-estimation for this bone with subsequent works (Kranioti & Michalodimitrakis, 2009; Reddy & Doshi, 2017), showing accuracies up to 92.9% and thus providing more evidence of the importance of other regions of the bone for the sexing of individuals. The ulna has also been the subject of research regarding its potential for the estimation of sex, with the caveat of the models being more population-specific than other bones, but achieving accuracies up to 84.2% of correct classifications (Cowal & Pastor, 2008). Finally, the research on the radius has shown good results, with achieved accuracies of up to 93.4% under cross-validation and 90% when tested on a separate sample (Curate et al., 2021).

The increased availability of data, along with ease of storage and free and open source computational resources has had an invaluable effect on the development and sharing of new methods (d'Oliveira Coelho et al., 2020; Klales, 2020; Santos et al., 2020).

2. Materials and Methods

2.1. Materials

A main sample from the Coimbra Identified Skeletal Collection (CISC, Coimbra, Portugal) was used to train and fit the models for sex estimation. Measurements of the left humerus, ulna and radius were collected from 280 adult individuals (140 males and 140 females), of Portuguese descent, with ages ranging from 19 to 89 years old (mean = 47.43). The data from the left radii measurements was taken from the original dataset from Neto (1957).

The same measurements were collected to form a test sample with 50 adult individuals (22 males and 28 females) from the 21st Century Identified Skeletal Collection (CEI/XXI, Laboratory of Forensic Anthropology, Coimbra, Portugal), with ages ranging from 62 to 96 years old (mean = 82.42), and was used to assess and test the methods developed.

2.1.1. Coimbra Identified Skeletal Collection

The Coimbra Identified Skeletal Collection (CISC), housed at the Department of Life Sciences (DCV) of the University of Coimbra is composed of skeletal remains from individuals exhumed from the *Conchada* cemetery in Coimbra that were born between the years of 1817 and 1924 and died between the years of 1904 and 1938.

A total of 505 complete skeletons, 266 male and 239 female individuals, make up the collection with 498 of the 505 individuals collected from the *Conchada* cemetery, while 7 are from dissected cadavers. Nearly all individuals are of Portuguese descent, while there are 9 with indicated places of birth in Africa, Spain, or Brazil. The 266 males have ages ranging from 7 to 96 years old, with a mean age of 43.44, while the female individuals' ages range from 7 to 90 years old with a mean value of 46.55.

A registry book accompanies the collection and has information pertaining to each of the individuals, such as name, sex, age, cause of death, etc (Cunha & Wasterlain, 2007). The completeness of the skeletons and the available information make this a very important collection in the context of forensic and biological anthropology.

2.1.2. 21st Century Identified Skeletal Collection

The 21st Century Identified Skeletal Collection, housed at the Laboratory of Forensic Anthropology, Department of Life Sciences (DCV) of the University of Coimbra, Portugal, is made up of skeletons from individuals from the late 20th/early 21st century. These were exhumed from the *Capuchos* cemetery in Santarém, the result of a protocol between the former Department of Anthropology, currently integrated within the DCV and the local authorities responsible for the cemetery. The skeletons were unclaimed by relatives and thus allowed to be donated to the University (Ferreira et al., 2014). As is the case with other identified collections, a plethora of information is already available about the individuals, such as sex, weight, osteometric data, etc (Ferreira et al., 2021).

Currently the collection is composed of 302 adult skeletons of both sexes, 162 female individuals with an age-at-death range between 28 and 101 years old, with a mean of 81.19 and a standard deviation of 12.89. There are less males, 140 individuals, and they have died younger, with age-at-death ranging from 25 to 96 years old, mean = 73.20 and sd = 15.61. All individuals died between 1982 and 2012 (Ferreira et al., 2021). The individuals in the collection are quite old, providing ample opportunity for the development and testing of methods in older adults (Ferreira et al., 2014).

Within the aforementioned collection, a subgroup of skeletons is being burned under controlled conditions with the intent to improve the knowledge on heat induced changes in human bones and teeth and the testing and developing of methods for assessing the biological profile of burnt skeletal remains (Ferreira et al., 2021).

2.2. Methods

2.2.1. Measurements

2.2.1.1. Humerus

A total of six measurements (taken in millimeters) were selected from the humerus, including the maximum length (HML), vertical diameter of the head (HVDH), horizontal diameter of the head (HVDH), minimum circumference (HMC), epicondylar breadth (HEB) and condilar breadth (HCB). Measurement definitions and descriptions are available in Table 1, and depicted in Figures 1.

Table 1 - Anthropological measurements of the humerus

Measurement	Definition	Reference
Maximum Length (HML)	The distance from the most superior point on the head of the humerus to the most inferior point on the trochlea. Instrument: *osteometric board*	(Langley et al., 2016): 74 #45)
Vertical Diameter of the Head (HVDH)	The distance between the most superior and inferior points on the border of the articular surface. Instrument: *sliding calliper*	(Langley et al., 2016: 74 #47)
Horizontal Diameter of the Head (HHDH)	The maximum breadth of the humeral head taken in the anterior-posterior direction on the articular surface. Instrument: *sliding calliper*	(Byrd & Adams, 2003) #42A)
Minimum Circumference (HMC)	The least circumference taken below the deltoid tuberosity. Instrument: *measuring tape*	
Epicondylar Breadth (HEB)	The distance from the most laterally protruding point on the lateral epicondyle to the corresponding projection on the medial epicondyle. Instrument: *sliding calliper*	(Langley et al., 2016: 74, #46)
Condilar Breadth (HCB)	The breadth of the capitulum and trochlea at the distal humerus. Instrument: *sliding calliper*	(Byrd and Adams, 2003 #41A)

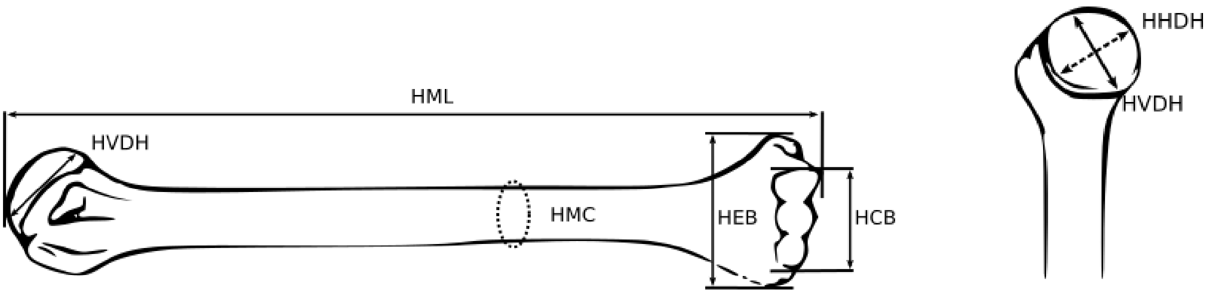

Figure 1- Osteometric dimensions of the humerus.

2.2.1.2. Ulna

A total of six measurements (taken in millimeters) were selected from the ulna, including the maximum length (UML), physiological length (UPL), minimum circumference (UMC), olecranon breadth (UOB), breadth of the semilunar notch (UBSN) and the distal articular width (UDAW). Measurement definitions and descriptions are available in Table 2, and depicted in Figure 2.

Table 2 - Anthropological measurements of the ulna

Measurement	Definition	Reference
Maximum Length (UML)	The distance between the most proximal point on the olecranon and the most distal point on the styloid process. Instrument: *osteometric board*	(Langley et al., 2016: 75, #54)
Physiological Length (UPL)	The distance between the deepest point on the articular surface of the coronoid process on the guiding ridge and the most inferior point on the distal articular surface of the ulna. Instrument: *spreading calliper*	(Langley et al., 2016: 76, #57)
Minimum Circumference (UMC)	The least circumference near the distal end of the bone. Instrument: *measuring tape*	(Langley et al., 2016: 76, #58)
Olecranon Breadth (UOB)	The maximum breadth of the olecranon process, taken perpendicular to the longitudinal axis of the semilunar notch. Instrument: *sliding calliper*	(Langley et al., 2016: 76, #59)
Breadth of the Semilunar Notch (UBSN)	This is a measure of only the distal surface of the semilunar notch (the base). Instrument: *sliding calliper*	(Byrd and Adams, 2003 #51C)
Distal Articular Width (UDAW)	The width of the distal articular surface taken in the anterior position. Instrument: *sliding calliper*	

Figure 2 - Osteometric measurements of the ulna.

2.2.1.3. Radius

A total of five measurements (taken in millimeters) were selected from the radius, including the maximum length (RML), minimum circumference of the distal diaphysis (RMCDD), head circumference (RHC), neck circumference (RNC) and the width of the distal epiphysis (RDEW). Measurement definitions and descriptions are available in Table 3, and depicted in Figure 3.

Table 3 - Anthropological measurements of the radius

Measurement	Definition	Reference
Maximum Length (RML)	The distance from the most superior point on the head of the radius to the apex of the styloid process. Instrument: *osteometric board*	
Minimum Circumference in the Distal Diaphysis (RMCDD)	The perimeter taken in the thinnest diaphyseal region located at the distal part of the bone. Instrument: *measuring tape*	
Head Circumference (RHC)	Maximum perimeter of the radius head. Instrument: *measuring tape*	(Martin, 1928 in Neto, 1957)
Neck Circumference (RNC)	Minimum perimeter of the neck, taken in the most strangled place of the neck. Instrument: *measuring tape*	
Distal Epiphysis Width (RDEW)	Maximum projective distance from the ulnar notch to the lateral aspect of the styloid process. Instrument: *sliding calliper*	

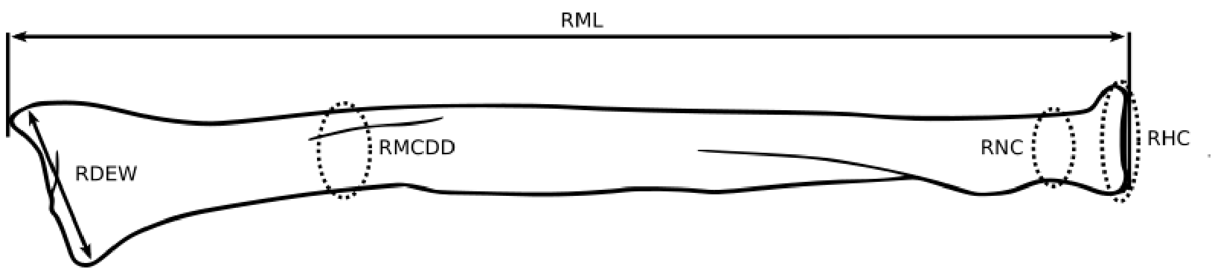

Figure 3 - Osteometric measurements of the radius. (Adapted from Langley et al., 2016).

2.2.2. Technical Error of Measurement

Inter and intra observer error were calculated using the Technical Error of Measurement (TEM) (Perini et al., 2005). The process includes the collection of measurements from the same 30 individuals at two different occasions, once at the beginning and, the second time, at the end of data collection. The absolute TEM (error expressed in millimeters) is calculated with the following equation:

$$Absolute\,TEM = \sqrt{\frac{\sum d_i^2}{2n}}$$

Where $\sum d_i^2$ is the summation of the differences between the 1st and 2nd measurements of each individual raised to the second power for a given variable and n is the number of individuals whose measurements were taken. When applied to the above equation we get the value, in millimeters, of the absolute error per measurement. To obtain the Relative TEM, the error expressed as a percentage, we follow with a second step:

$$Relative\,TEM = \frac{TEM}{VAV} \times 100$$

where the previous calculated value of TEM is divided by the *Variable average value* VAV, which we get by adding the values of the first and second measurements performed for each individual and dividing it by two. This is repeated for each individual and all the mean values are finally added and divided by the total number of individuals.

Only measurements with Relative TEM lower than 5% were considered for analysis. For each analysis only individuals with a complete set of measurements were considered.

2.2.3. Statistical Analysis

All standard descriptive statistics were calculated. Group means, standard deviation (SD) and 95% confidence intervals (95% CI) for the mean of the continuous variables. Normal distributions of the variables were ascertained with the values of skewness and kurtosis and homoscedasticity was assessed by a Levene's test. Independent samples t-test were calculated to evaluate the null hypothesis that the means of every measurement in males and females were identical. The sexual dimorphism index (SDI) was also estimated for both the training and testing samples, following:

$$SDI = \frac{\bar{x}_m - \bar{x}_f}{\bar{x}_m} \times 100$$

where $\bar{x}_m$ and $\bar{x}_f$ are the mean values of any given measurement for males and females, respectively.

Sex estimation models were generated with a classical (logistic regression [LR]) and two machine-learning (support vector machines [SVM] and decision trees [C5.0]) algorithms. Logistic regression, a non-parametric classification method, models the probability of two mutually exclusive classes of a dependent variable (Klales et al., 2020). Unlike LDA (Linear discriminant analysis), an other widely used statistical method for the generation of sex estimation functions, LR provides the p-value that is first used to allocate an unkown individual and is second used to make a probabilistic statement about the likelihood of a correct allocation for any given case (Bartholdy et al., 2020; Nikita & Nikitas, 2020). A support vector machine (SVM) is a machine learning algorithm used for two-group classification problems, that builds a model on a dataset of labeled data and then categorizes new examples. For this work our categories are either M (for male) or F (for female), and the support vector machine outputs a hyperplane (decision boundary) between the two categories and then the new data falls to one of the two sides of the boundary (Kuhn & Johnson, 2013). The C5.0 algorithm was used to generate decision trees and provide sectioning points for the sex estimation for each variable. This algorithm builds and improves on the widely used C4.5 (Quinlan, 1993). Like C4.5 it has a tree-based version and shares much of its core functionality with its predecessor. However, the improvements in reduced error rates, smaller and less complex trees and lower execution times (up to 6x faster for this particular case) made it the better option (Kuhn & Johnson, 2013).

Finally all models were tested with a ten-fold cross-validation method. It partitions the training data into 10 equal-sized chunks, then trains 10 distinct models. Each model is trained on the union of 9 chunks of the data, and tested on the held out chunk. The average performance of these 10 classifiers stands in as the presumed accuracy for the full model (Skiena, 2017).

The goodness of fit of the provisional and cross-validated models, as well as the testing sample, was evaluated through the overall accuracy (a measure of total agreement between the real and the predicted sex), the proportion of females and males properly grouped and Cohen's Kappa (Larose, 2015).

All statistical analyses were performed with the free and open source R programming language (R Core Team & Others, 2013) in the R Studio environment (RStudio Team, 2020) for its ease of use and impact on reproducible research (Glennie, 2021). A combination of R packages were used but most importantly the *caret* package along with all its utilities (Kuhn, 2008; Kuhn & Johnson, 2013).

3. Results

3.1. Technical Error of Measurement

3.1.1. Intra-observer error

Intra-observer error was assessed with the collection of repeated measurements from the left humerus and left ulna of the same 30 individuals from the Coimbra Identified Skeletal Collection at two different occasions, once at the beginning of data collection and, the second time, approximately 4 weeks after. Error measurements are indicated in Table 4. Full measurement data is available in Tables A1 and A2 of the annexes.

For the humerus measurements, *TEM* was the lowest for the vertical head diameter (0.22 mm) and the highest for the maximum length (1.36 mm), while the *Relative TEM* was never above 1.6%, providing confidence in the measurements collected. Ulnar measurements had the lowest *TEM* for the distal articular width (close to zero) and the highest for maximum length (1.41 mm), while the *Relative TEM* was kept below 3.5%. All values for the *Relative TEM* are below the 5% proving good reproducibility.

Table 4 - Intra observer error

Measurement	TEM (mm)	VAV	Relative TEM (%)
Maximum Length (HML)	1.36	306.18	0.44
Vertical Diameter of the Head (HVDH)	0.22	42.04	0.53
Horizontal Diameter of the Head (HHDH)	0.56	38.98	1.44
Minimum Circumference (HMC)	0.93	56.26	1.57
Epicondylar Breadth (HEB)	0.22	57.07	0.38
Condilar Breadth (HCB)	0.56	41.43	1.34
Maximum Length (UML)	1.41	241.22	0.58
Physiological Length (UPL)	0.32	213.46	0.15
Minimum Circumference (UMC)	1.16	33.82	3.44
Olecranon Breadth (UOB)	0.32	23.37	1.38
Breadth of the Semilunar Notch (UBSN)	0.59	22.46	2.64
Distal Articular Width (UDAW)	<0.01	15.53	<0.01

Legend: TEM = technical error of measurement, VAV = variable average value, rTEM = relative technical error of measurement

3.1.2. Inter-observer error

Inter-observer error was assessed with a comparison between measurement data from the left radius, collected alongside the above mentioned humerus and ulna data, of 30 individuals of the Coimbra Identified Skeletal Collection and the same individuals from the data collected by Neto (1957). The lowest *TEM* was observed for the neck circumference (0.13 mm) and the highest for head circumference (1.42 mm), while *Relative TEM* values were all below 4.2%. RDEW is taken in a very sensible location of the radius and wear resulting from the handling of the bones might be the reason for the higher relative error observed.

Error measurements are indicated in Table 5. Full measurement data is available in Tables B1 and B2 of the annexes.

Table 5 - Inter observer error

Measurement	TEM (mm)	VAV	Relative TEM (%)
Maximum Length (RML)	1.23	224.64	0.56
Minimum Circumference in the Distal Diaphysis (RMCDD)	0.39	37.48	1.03
Head Circumference (RHC)	1.42	64.07	2.22
Neck Circumference (RNC)	0.13	43.03	0.30
Distal Epiphysis Width (RDEW)	1.28	30.92	4.13

Legend: TEM = technical error of measurement, VAV = variable average value, rTEM = relative technical error of measurement

3.2. Descriptive Statistics and Sectioning Points

3.2.1. Humerus

All metric variables are normally distributed. Summary descriptive statistics for the training (Coimbra Identified Skeletal Collection) and test (21st Century Identified Skeletal Collection) samples are shown in Tables 6 and 7. Measurements from 132 left humerus (121 female and 111 male) were selected from the training samples and 35 (19 female and 16 male) from the holdout sample. An index of sexual dimorphism (SDI) is also shown, varying between 7.9 (maximum length) and 14.6 (condilar breadth) in the CISC sample, and between 6.3 (maximum length) and 11.2 (minimum circumference) in the CEI/XXI sample.

Table 6 - Descriptive statistics for the left humerus in both sexes; Coimbra Identified Skeletal Collection (CISC)

Measurement	Females				Males				SDI	t-test	sig.
	Mean	SD	95%CI	N	Mean	SD	95%CI	N			
HML	290.4	16.4	281.5-293.4	121	315.4	17.8	312.1-318.8	111	7.9	-11.09	<0.001
HMC	54.5	3.7	53.8-55.2	121	63.4	3.9	62.7-64.2	111	14.0	-17.76	<0.001
HVDH	39.1	2.1	38.7-39.5	121	45.3	2.3	44.8-45.7	111	13.7	-21.0	<0.001
HHDH	35.9	1.9	35.5-36.2	121	41.7	2.1	41.3-42.1	111	13.9	-22.6	<0.001
HEB	52.5	2.1	51.9-53.0	121	61.4	3.3	60.8-62.1	111	14.5	-21.0	<0.001
HCB	38.0	3.2	37.7-38.4	121	44.5	2.6	44.0-45.0	111	14.6	-20.7	<0.001

Legend: SD = standard deviation, 95%CI = confidence intervals of the mean at 95%, N = number of individuals, SDI = sexual dimorphism index

A sectioning point for each metric variable was computed with a decision-tree algorithm (C5.0). Performance for all sectioning points are summarized in Table 8. Accuracy under cross-validation varies from 79.3% to 94.4%, while bias, or systematic error (the absolute difference between the percentage of correctly classified females and males), varies from 0.4% to 5.9%. The variables with the best overall performance under cross validation were the horizontal diameter of the head (accuracy: 94.4%, bias:

2.2%, kappa: 0.887, sectioning point: 38.8 mm) and the condilar breadth (accuracy: 94.4%, bias: 0.4%, kappa: 0.888, sectioning point: 40.7 mm).

Table 7 - Descriptive statistics for the left humerus in both sexes; 21st Century Identified Skeletal Collection (CEI/XXI)

Measurement	Females				Males				SDI	t-test	sig.
	Mean	SD	95%CI	N	Mean	SD	95%CI	N			
HML	291.7	17.2	283.4-300.0	19	311.3	14.6	303.5-319.1	16	6.3	-3.6	<0.001
HMC	56.9	4.3	54.8-59.0	19	64.1	4.2	61.8-66.3	16	11.2	-5.0	<0.001
HVDH	41.2	2.8	39.8-42.6	19	45.5	2.7	44.1-46.9	16	9.5	-4.6	<0.001
HHDH	38.0	2.7	36.7-39.3	19	42.3	2.1	41.2-43.4	16	10.2	-5.3	<0.001
HEB	55.1	4.4	53.0-57.2	19	61.0	4.5	58.6-63.4	16	9.7	-3.9	<0.001
HCB	39.2	2.2	38.1-40.2	19	44.0	2.3	42.8-45.3	16	10.9	-6.3	<0.001

Legend: SD = standard deviation, 95%CI = confidence intervals of the mean at 95%, N = number of individuals, SDI = sexual dimorphism index

In the test sample, overall accuracy ranged from 68.4% to 77.1%, with bias ranging from 7.6% to 30.6%. The best predictors of sex for the test sample, accounting for accuracy and bias, were the condilar breadth (accuracy: 77.1%, bias: 7.6%, kappa: 0.544, sectioning point: 40.7 mm) and epicondylar breadth (accuracy: 77.1%, bias: 19.1%, kappa: 0.548, sectioning point: 56.7 mm).

Table 8 - Sectioning points and goodness of fit for the humeral measurements

Meas.	Sectioning Point	Training sample				Cross Validation				Test Sample			
		Accuracy	Females	Males	Kappa	Accuracy	Females	Males	Kappa	Accuracy	Females	Males	Kappa
HML	301.5	0.815	0.820	0.810	0.629	0.793	0.793	0.784	0.587	0.686	0.579	0.813	0.382
HMDC	58.0	0.888	0.860	0.919	0.776	0.888	0.860	0.919	0.776	0.771	0.632	0.938	0.553
HVDH	41.7	0.935	0.926	0.946	0.871	0.931	0.926	0.937	0.861	0.714	0.579	0.875	0.441
HHDH	38.8	0.948	0.959	0.937	0.896	0.944	0.959	0.937	0.887	0.771	0.632	0.938	0.553
HEB	56.7	0.931	0.926	0.937	0.862	0.931	0.926	0.937	0.862	0.771	0.684	0.875	0.548
HCB	40.7	0.948	0.950	0.946	0.896	0.944	0.942	0.946	0.888	0.771	0.737	0.813	0.544

Legend: Accuracy = proportion of correctly assigned individuals, Females = proportion of correct assigned females; Males = proportion of correctly assigned individuals, Kappa = Cohen’s Kappa

3.2.2. Ulna

Complete measurements of the left ulna were taken from a total of 256 individuals from the Coimbra Identified Skeletal Collection, 129 female and 127 male. From the holdout samples 38 left ulnas were selected, 20 female and 18 male.

All metric variables are normally distributed. Descriptive statistics for both sexes along with the sexual dimorphism index (SDI) and independent t-test results are shown in Table 9. SDI ranges from 10.0 to 16.8 (maximum length and breadth of the semilunar notch, respectively).

Table 9 - Descriptive statistics for the ulna in both sexes; Coimbra Identified Skeletal Collection

Measurement	Females				Males				SDI	t-test	sig.
	Mean	SD	95%CI	N	Mean	SD	95%CI	N			
UML	229.7	14.7	227.2-232.3	129	255.2	13.6	252.8-257.6	127	10.0	-14.4	<0.001
UPL	202.5	10.4	200.7-204.3	129	225.3	13.4	222.9-227.6	127	10.1	-15.2	<0.001
UMC	31.5	2.9	31.0-32.0	129	36.5	3.4	35.9-37.1	127	13.7	-12.6	<0.001
UOB	21.2	1.8	20.8-21.5	129	24.9	1.9	24.6-25.2	127	14.9	-16.1	<0.001
UBSN	19.8	1.6	19.6-20.1	129	23.8	2.0	23.4-24.1	127	16.8	-17.0	<0.001
UDAW	14.3	1.4	14.1-14.6	129	16.7	1.6	16.4-16.9	127	14.4	-12.3	<0.001

Legend: SD = standard deviation, 95%CI = confidence intervals of the mean at 95%, N = number of individuals, SDI = sexual dimorphism index

A sectioning point for each metric variable was computed with a decision-tree algorithm (C5.0). Performance for all sectioning points are summarized in Table 11. Accuracy under cross-validation varies from 75.4% to 87.1%, while bias, or systematic error (the absolute difference between the percentage of correctly classified females and males), varies from 0.6 to 34.5%. The variables with the best overall performance under cross validation were the breadth of the semilunar notch (accuracy: 87.1%, bias: 0.6%%, kappa: 0.751, sectioning point: 21.4 mm) and the olecranon breadth (accuracy: 85.9%, bias: 6.4%, kappa: 0.719, sectioning point: 23.4 mm).

Table 10 - Descriptive statistics for the ulna in both sexes; 21st Century Identified Skeletal Collection

Measurement	Females				Males				SDI	t-test	sig.
	Mean	SD	95%CI	N	Mean	SD	95%CI	N			
UML	233.0	14.0	226.4-239.5	20	254.3	16.1	246.3-262.4	18	8.4	-4.34	<0.001
UPL	206.4	13.0	200.3-212.5	20	223.8	15.6	216.0-231.5	18	7.8	-3.7	<0.001
UMC	32.4	2.7	31.1-33.7	20	37.1	2.5	35.9-38.3	18	12.7	-5.6	<0.001
UOB	21.7	2.0	20.8-22.7	20	25.3	1.4	24.6-26.0	18	14.2	-6.5	<0.001
UBSN	20.7	1.8	19.9-21.6	20	24.5	1.8	23.6-25.4	18	15.5	-6.4	<0.001
UDAW	15.5	2.0	14.6-16.4	20	18.0	1.3	17.4-18.7	18	13.9	-4.7	<0.001

Legend: SD = standard deviation, 95%CI = confidence intervals of the mean at 95%, N = number of individuals, SDI = sexual dimorphism index

In the test sample, overall accuracy ranged from 60.5% to 79.0%, with bias ranging from 7.6% to 75%. The best predictors of sex for the test sample, accounting for accuracy and bias, were the maximum length (accuracy: 77.1%, bias: 7.6%, kappa: 0.544, sectioning point: 246.5 mm) and physiological length (accuracy: 76.3%, bias: 18.3%, kappa: 0.521, sectioning point: 218.0 mm). By far the worst indicator in the test sample was the distal articular width with an accuracy of 60.5% with a bias of 75%.

Table 11 - Sectioning points and goodness of fit for the ulna

Meas.	Sectioning Point	Training sample				Cross Validation				Test Sample			
		Accuracy	Females	Males	Kappa	Accuracy	Females	Males	Kappa	Accuracy	Females	Males	Kappa
UML	246.5	0.859	0.954	0.764	0.718	0.852	0.938	0.764	0.704	0.771	0.737	0.813	0.544
UPL	218.0	0.836	0.946	0.724	0.671	0.824	0.845	0.803	0.648	0.763	0.850	0.667	0.521
UMC	32.5	0.801	0.690	0.913	0.602	0.801	0.690	0.913	0.602	0.737	0.550	0.944	0.484
UOB	23.4	0.867	0.853	0.882	0.734	0.859	0.891	0.827	0.719	0.790	0.650	0.944	0.585
UBSN	21.4	0.879	0.868	0.890	0.758	0.871	0.868	0.874	0.741	0.790	0.650	0.944	0.585
UDAW	14.2	0.750	0.535	0.969	0.502	0.754	0.581	0.929	0.510	0.605	0.250	1.000	0.240

Legend: Accuracy = proportion of correctly assigned individuals, Females = proportion of correct assigned females; Males = proportion of correctly assigned individuals, Kappa = Cohen's Kappa

3.2.3. Radius

Complete measurements of the left radius were selected from a total of 231 individuals from the Coimbra Identified Skeletal Collection, 106 female and 125 male, and 34 (20 female and 14 male) from the 21st Century Identified Collection .

All metric variables are normally distributed. Descriptive statistics for both sexes along with the sexual dimorphism index (SDI) and independent t-test results are shown in Tables 12 and 13. SDI ranges from 11.0 to 16.3 (maximum length and minimum circumference of the distal diaphysis, respectively).

Table 12 - Descriptive statistics for the radius in both sexes; Coimbra Identified Skeletal Collection

Measurement	Females				Males				SDI	t-test	sig.
	Mean	SD	95%CI	N	Mean	SD	95%CI	N			
RML	210.2	10.8	208.2-212.3	106	236.3	12.8	234.0-238.5	125	11.0	-16.8	<0.001
RMCDD	34.0	2.8	33.5-34.1	106	40.6	3.0	40.0-41.1	125	16.3	-17.4	<0.001
RHC	59.3	3.8	58.6-60.0	106	69.4	4.3	68.6-70.2	125	14.6	-19.1	<0.001
RNC	39.7	3.6	39.0-40.4	106	46.0	3.9	45.4-46.7	125	13.7	-12.9	<0.001
RDEW	28.8	1.5	28.5-29.1	106	33.4	2.1	33.0-33.8	125	13.8	-19.4	<0.001

Legend: SD = standard deviation, 95%CI = confidence intervals of the mean at 95%, N = number of individuals, SDI = sexual dimorphism index

Table 13 - Descriptive statistics for the radius in both sexes; 21st Century Identified Skeletal Collection

Measurement	Females				Males				SDI	t-test	sig.
	Mean	SD	95%CI	N	Mean	SD	95%CI	N			
RML	222.4	27.4	209.6-235.3	20	233.8	14.8	225.2-242.3	14	4.9	-1.5	<0.01
RMCDD	36.0	3.1	34.5-37.4	20	42.2	3.0	40.5-43.9	14	14.7	-5.9	<0.001
RHC	60.3	3.5	58.7-62.0	20	68.3	4.7	65.6-71.0	14	11.7	-5.4	<0.001
RNC	40.5	3.5	38.9-42.1	20	46.9	3.1	45.2-48.7	14	13.6	-5.7	<0.001
RDEW	29.5	1.9	28.6-30.4	20	33.3	1.9	32.3-34.4	14	11.4	-5.8	<0.001

Legend: SD = standard deviation, 95%CI = confidence intervals of the mean at 95%, N = number of individuals, SDI = sexual dimorphism index

A sectioning point for each metric variable was computed with a decision-tree algorithm (C5.0). Performance for all sectioning points are summarized in Table 14. Accuracy under cross-validation varies from 80.5% to 90.5%%, while bias, or systematic error (the absolute difference between the percentage of correctly classified females and males), varies from 1 to 11.6%%. The variables with the best overall performance under cross validation were the maximum length (accuracy: 88.2%, bias: 1%, kappa: 0.703, sectioning point: 221.5 mm) and the head circumference (accuracy: 90.5%, bias: 3.7%, kappa: 0.809, sectioning point: 64.0 mm).

In the test sample, overall accuracy ranged from 70.6% to 82.4%, with bias ranging from 5.7% to 32.9%. The best predictors of sex for the test sample, accounting for accuracy and bias, were the head circumference (accuracy: 82.4%, bias: 5.7%, kappa: 0.643, sectioning point: 64 mm), nech circumference (accuracy: 82.4%, bias: 17.9%, kappa: 0.651, sectioning point: 43 mm) and distal epypsis width (accuracy: 82.4%, bias: 17.9%, kappa: 0.651, sectioning point: 30.5 mm).

Table 14 - Sectioning points and goodness of fit for the radius

Meas.	Sectioning Point	Training sample				Cross Validation				Test Sample			
		Accuracy	Females	Males	Kappa	Accuracy	Females	Males	Kappa	Accuracy	Females	Males	Kappa
RML	221.5	0.883	0.868	0.896	0.765	0.852	0.858	0.848	0.703	0.706	0.600	0.857	0.430
RMDDC	36.0	0.887	0.859	0.912	0.773	0.887	0.858	0.912	0.773	0.735	0.600	0.929	0.492
RHC	64.0	0.905	0.925	0.888	0.809	0.905	0.925	0.888	0.809	0.824	0.800	0.857	0.643
RNC	43.0	0.814	0.859	0.776	0.629	0.805	0.868	0.752	0.613	0.824	0.750	0.929	0.651
RDEW	30.5	0.909	0.877	0.936	0.816	0.904	0.868	0.936	0.806	0.824	0.750	0.929	0.651

Legend: Accuracy = proportion of correctly assigned individuals, Females = proportion of correct assigned females; Males = proportion of correctly assigned individuals, Kappa = Cohen’s Kappa

3.3. Multivariate models

Multivariate models for sex estimation were generated with logistic regression (LR) and support vector machines (SVM) (Table 15). The goodness of fit for the suggested models are shown in Tables 16 and 17. Logistic regression models's accuracy ranges from 84.4% to 96.7% (bias: 0.3% to 5.6%) for cross validation and 74.3% to 86.4 (bias: 12.9% to 35.9%) in the test sample, while the SVM models yield accuracies ranging from 90.2% to 96.2% (bias 0.3% to 13.4%) under cross-validation and 77.1% to 86.4% (bias: 12.9 to 30.8%) in the test sample.

Table 15 - Logistic Regression (LR) classification functions for the different multivariate models

Model	Classification function
LR_H1	$(1.0598 \times hhdh) + (0.6616 \times heb) - 78.6127$
LR_U1	$(0.91312 \times ubsn) + (0.08831 \times uml) - 41.09560$
LR_U2	$(0.71146 \times ubsn) + (0.07679 \times uml) + (0.33344 \times uob) - 41.61076$
LR_U3	$(0.82916 \times ubsn) + (0.07617 \times uml) + (0.17234 \times umc) - 42.19127$
LR_R1	$(0.3725 \times rhc) + (0.4488 \times rmddc) - 40.3878$
LR_R2	$(0.2921 \times rhc) + (1.0871 \times rdew) - 51.9642$
LR_R3	$(0.4406 \times rmddc) + (1.1125 \times rdew) - 50.4790$
LR_HU1	$(1.4021 \times hhdh) + (0.5999 \times ubsn) - 67.2493$
LR_HR1	$(1.329 \times hhdh) + (1.067 \times rdew) - 84.273$
LR_HR2	$(1.5631 \times hhdh) + (0.7454 \times rmddc) - 87.8670$
LR_HR3	$(1.5052 \times hhdh) + (0.1823 \times rnc) - 65.9151$
LR_UR1	$(0.062 \times uml) + (1.565 \times rdew) - 63.329$
LR_UR2	$(1.4521 \times rdew) + (0.2664 \times umc) - 53.8541$
LR_HUR1	$(1.2524 \times hhdh) + (0.2062 \times ubsn) + (0.9116 \times rdew) - 80.9951$

For the humerus, a single model was selected using the horizontal diameter of the head (HHDH) and the epicondylar breadth (HEB) with the LR model yielding an accuracy of 95.2% (bias of 2.2%) under cross validation, opposingly it was the least performant model in the test sample with an accuracy under cross validation of 74.3% and bias of 35.9%. The corresponding SVM model had an accuracy of 96.1% (bias 4.6%) under cross-validation and 77.1% (bias 30.6%) in the test sample.

Table 16 - Logistic Regression (LR) classification accuracy with the different multivariate models

Model	Training sample				Cross Validation				Test Sample			
	Accuracy	Females	Males	Kappa	Accuracy	Females	Males	Kappa	Accuracy	Females	Males	Kappa
LR_H1	0.957	0.950	0.964	0.914	0.952	0.942	0.964	0.905	0.743	0.579	0.938	0.499
LR_U1	0.895	0.923	0.866	0.789	0.894	0.922	0.866	0.789	0.816	0.700	0.944	0.636
LR_U2	0.910	0.923	0.898	0.820	0.902	0.922	0.882	0.804	0.816	0.700	0.944	0.636
LR_U3	0.910	0.923	0.898	0.820	0.902	0.922	0.882	0.804	0.790	0.650	0.944	0.585
LR_R1	0.918	0.906	0.928	0.834	0.926	0.925	0.928	0.852	0.853	0.800	0.929	0.706
LR_R2	0.918	0.925	0.912	0.835	0.922	0.925	0.920	0.843	0.853	0.800	0.929	0.706
LR_R3	0.922	0.915	0.928	0.843	0.922	0.915	0.928	0.842	0.853	0.800	0.929	0.706
LR_HU1	0.968	0.966	0.969	0.935	0.967	0.966	0.969	0.934	0.818	0.692	1.000	0.648
LR_HR1	0.957	0.955	0.958	0.913	0.962	0.955	0.969	0.924	0.818	0.692	1.000	0.648
LR_HR2	0.962	0.966	0.958	0.924	0.957	0.955	0.958	0.914	0.818	0.692	1.000	0.648
LR_HR3	0.951	0.955	0.948	0.903	0.951	0.955	0.948	0.902	0.818	0.692	1.000	0.648
LR_UR1	0.930	0.932	0.927	0.859	0.929	0.933	0.927	0.858	0.818	0.692	1.000	0.648
LR_UR2	0.930	0.933	0.927	0.859	0.929	0.921	0.938	0.857	0.864	0.769	1.000	0.732
LR_HUR1	0.962	0.955	0.969	0.924	0.956	0.955	0.958	0.913	0.818	0.692	1.000	0.648

Legend: Accuracy = proportion of correctly assigned individuals, Females = proportion of correct assigned females; Males = proportion of correctly assigned individuals, Kappa = Cohen's Kappa

Three models were selected for the ulna. LR_U1 used the breadth of the semilunar notch (UBSN) along with the maximum length (UML) for an accuracy of 89.4% and bias of 5.6% under cross validation. Both the second and third models included the aforementioned measurements with the addition of the olecranon breadth

(UOB) and minimum circumference (UMC), respectively. Both had accuracies under cross-validation of 90.2% and bias of 4%. Testing in the holdout sample revealed accuracies of 81.6% for all the models, however bias was 24.4% for the first model and 14.4% for the others.

SVM models for the ulna show accuracies from 90.2% to 91.4% (bias: 3.2% to 13.4%), while for the holdout sample accuracies yield values from 79% to 84.2% with bias ranging from 19.4% to 29.4%.

For the radius three models were selected with very similar accuracies both under cross-validation and on the holdout sample. LR_R1 made use of the measurements of the head circumference (RHC) and the minimum distal diaphysis circumference (RMDDC) and showed an accuracy of 92.6% and bias of 0.3% under cross-validation. For LR_R2 rmddc was replaced with the distal epiphysis width (RDEW) and had an accuracy of 92% and bias 0.5% under cross-validation. The final model included rmddc and rdew for an accuracy under cross-validation of 92.8% and bias of 1.3%. All three models revealed accuracies of 85.3% and bias of 12.9% when tested on the holdout sample. SVM models had very similar accuracies under cross validation and on the test sample.

A single model incorporating both humerus and ulna measurements was selected with the horizontal diameter of the head of the humerus (HHDH) and the breadth of the semilunar notch of the ulna. Accuracies were 96.2% (bias: 1.4%) and 81.8% (bias: 30,8%) for cross-validation and the holdout sample, respectively. The SVM model yielded very similar accuracies. LR models involving measurements from the humerus and radius yielded accuracies ranging from 95.6% to 96.2% (bias: 0.3% to 1.4%) in cross-validation and 81.8% (bias: 30.8%) on the test sample. SVM counterparts had the same accuracy and bias values.

All (LR and SVM) ulna and radius joint models had an accuracy of 92.9% (bias: 0.6% and 1.7%) under cross-validation and 81.8% (bias: 30.8%) and 86.4% (bias: 23.1%) while tested on the holdout sample.

Finally, a single model including measurements from all bones was designed, HHDH along with UBSN and RDEW were selected. Accuracy under cross-validation for the LR model was 95.6 (0.3%) and 96.2% (bias: 0.8%) for the SVM model. Both methods performed equally on the test sample, accuracy: 81.8% and bias 30.8%.

Table 17 - Support vector machine (SVM) classification accuracy with the different multivariate models

Model	Training sample				Cross Validation				Test Sample			
	Accuracy	Females	Males	Kappa	Accuracy	Females	Males	Kappa	Accuracy	Females	Males	Kappa
SVM_H1	0.961	0.984	0.937	0.922	0.961	0.983	0.937	0.921	0.771	0.632	0.938	0.553
SVM_U1	0.898	0.961	0.835	0.797	0.902	0.969	0.835	0.804	0.842	0.750	0.944	0.689
SVM_U2	0.914	0.930	0.898	0.828	0.914	0.930	0.898	0.828	0.816	0.700	0.944	0.636
SVM_U3	0.906	0.923	0.890	0.812	0.906	0.922	0.890	0.812	0.790	0.650	0.944	0.585
SVM_R1	0.926	0.925	0.928	0.852	0.931	0.934	0.928	0.861	0.853	0.800	0.929	0.706
SVM_R2	0.913	0.934	0.896	0.826	0.922	0.934	0.912	0.843	0.853	0.800	0.929	0.706
SVM_R3	0.922	0.915	0.928	0.843	0.922	0.915	0.928	0.842	0.853	0.800	0.929	0.706
SVM_HU1	0.962	0.955	0.969	0.924	0.962	0.955	0.969	0.923	0.818	0.692	1.000	0.648
SVM_HR1	0.957	0.955	0.958	0.913	0.956	0.955	0.958	0.913	0.818	0.692	1.000	0.648
SVM_HR2	0.962	0.966	0.958	0.924	0.962	0.966	0.958	0.925	0.818	0.692	1.000	0.648
SVM_HR3	0.957	0.966	0.948	0.914	0.957	0.966	0.948	0.913	0.818	0.692	1.000	0.648
SVM_UR1	0.930	0.921	0.938	0.859	0.929	0.921	0.928	0.857	0.818	0.692	1.000	0.648
SVM_UR2	0.930	0.933	0.927	0.859	0.929	0.921	0.938	0.857	0.864	0.769	1.000	0.732
SVM_HUR1	0.962	0.966	0.958	0.924	0.962	0.966	0.958	0.924	0.818	0.692	1.000	0.648

Legend: Accuracy = proportion of correctly assigned individuals, Females = proportion of correct assigned females; Males = proportion of correctly assigned individuals, Kappa = Cohen's Kappa

4. Discussion

Sex estimation of human skeletal remains is at the foreground of both archeological and forensic contexts. Although the pelvis is the most dimorphic region of the skeleton (Spradley & Jantz, 2011), most times human remains are incomplete or fragmented and other parts of the skeleton must be employed to assess biological sex. Following the pelvis, long bones and the cranium also provide avenues for accurate sex determination. The humerus, ulna and radius have been used individually or in tandem to develop methods for sex estimation with very good results (Albanese, 2013; Bidmos & Mazengenya, 2021; Cowal & Pastor, 2008; Curate et al., 2021; Lee et al., 2015; Reddy & Doshi, 2017; Wasterlain, 2000).

For this particular work, the humerus, ulna and radius were selected and a set of univariate and multivariate models were designed employing two Portuguese reference samples.

Firstly, let's look at the two reference samples and the differences between the sexes are evident across the complete set of measurements. The sexual dimorphism index (SDI) is smaller in the 21st Century Identified Skeletal Collection (Ferreira et al., 2014, 2021), the test sample, with the exception of the radial neck circumference (RNC) that has a similar value. While SDI provides an incomplete measure of sexual dimorphism, because it is based on sample means and does not take into account their variance, it is important to note because while mean values for the males appear to maintain similar levels between the samples, the trend for female individuals in the CEI/XXI shows a slight increase in most metric dimensions of all bones. The 21st Century Identified Skeletal Collection is chronologically a later sample and we must take into account the efects of secular trends and its effects on the metric variation of long bones and the effectiveness of sex estimation models (Gonçalves, 2014; Langley & Jantz, 2020). Looking at the mean ages of the samples we observe that the testing sample, CEI/XXI, has a mean age of 82.42 while the training sample shows a mean age of 47.43. This is a notable difference, however looking into the influence of age on the values for the measurements via statistical correlation tests, I found no evidence that the age of the individual had a significant influence in the values obtained for each measurement.

Univariate models for the humerus show good accuracy under cross-validation (between 79.3% and 94.4%) and low bias (0.4% to 5.9%), not unexpectedly the most accurate measurements were from the head and distal epiphysis (Wasterlain, 2000). Sectioning points for the ulna had accuracies under 87.1% and a very high bias for most measurements, contrary to both the humerus' and radius' univariate models where misclassifications alternated between males and females depending on the measurement. Finally the radius models yielded high accuracies (80.5% to 90.5%) and low bias (1% to 11.6%) under cross-validation, values in line with recent research on the same reference sample (Curate et al., 2021). Looking now at how all the multivariate models performed on the holdout sample we start to observe the high values of bias (up to 75%) that result from the excess of misclassified females in all but one of variables of the ulna (UPL). This pattern follows previously observed results in tarsal and femoral dimensions of Portuguese samples, in which systematic error was mostly a function of misclassified females - and might reflect changes caused by secular changes in the long bones of the upper limbs in the CISC and CEI/XXI reference samples (Curate et al., 2017; Navega et al., 2015). Now, secular trends in sexual dimorphism are not the only factor associated with changes in bone dimensions, genetics, hormonal status, nutritional disorders, and socioeconomic status, among others, can also play a large role (Albanese, 2010).

The best sectioning points for the humerus were the horizontal diameter of the head (38.8mm) along with condylar breadth (40.7mm) with similar accuracies both under cross-validation and on the test sample. The breadth of the semilunar notch (21.4mm) was the most accurate variable for the univariate models of the ulna and the head circumference of the radius yielded the best results for the third bone analysed in this work. As indicated above, the only variable with a relative poor performance was the distal articular width of the ulna.

Multivariable models yielded, expectedly, better accuracy and lower bias than the univariate models. Comparing the LR and SVM models, the performance metrics are very similar - with SVM-based models performing marginally better with higher values of accuracy and kappa, but a slight increase in bias.

The models were selected through stepwise selection, meaning that each variable was added or removed as it had a positive or negative (improving or not) effect on the final model. In the case of the single model for the humerus, any additional variable to those selected did not improve the outcomes and thus there was no need to add complexity for a poorer outcome. Moving into the ulna and the radius we have three models for each bone, in which a new variable is added or replaced on the previous model. This arrangement of different variables bolsters the models to respond and be applicable in contexts where skeletal remains or individual bones are incomplete or damaged.

Models for each individual bone performed relatively well under cross-validation with accuracies ranging from 89.4% to 95.2%, with the models for the ulna yielding the lowest accuracy results. The highest performing model was the one selected for the humerus with an accuracy under cross-validation of 95.2% but, curiously, this was also the worst performing model when tested on the holdout sample with an accuracy of 74.3% (bias: 35.9%). The models for the radius performed the best on the test sample out of all individual models.

Going into models mixing variables from different bones, the highest performers under cross-validation were HU1 (humerus and ulna variables), HR1 (humerus and radius variables) and HR2 and HUR1 (humerus, ulna and radius variables) for both Logistic Regression and Support Vector Machines models. The differences in cross-validated performance between the models is not too significant, but looking at the accuracies when tested on the holdout sample, UR2, which was the least accurate under cross-validation (accuracy: 92.9%), shows the highest accuracy (86.4%). Results for models performance were as anticipated from results from similar works from the same skeletal regions (Albanese, 2013; Curate et al., 2021).

5. Conclusion

The goals of this work were to develop univariate and multivariate models for sex estimation relying on mesurements from the long bones of the upper limbs, also look into how dimorphic these measurements are, and finally how these models perform when used to estimate the sex of a diferent population to that enlisted on the development of said models along with possible effects of secular trends on its uses. I feel those objectives were accomplished, with some important takeaways.

Not unlike other models, the performance under cross validation was quite good. Correct allocations of >90% are very important when estimating sex and developing biological profiles in forensic or archaeological settings. Looking at the performance when used on the test sample we start to observe some problems that must be taken into account when using the final equations. As we look at the values of the SDI for the two reference samples, the effects of secular trends start to be visible. On the latter sample (CEI/XXI) the values of the index for sexual dimorphism decrease, with mostly the female individuals being "bigger". This translates in the reduced accuracy and very high bias for the models with the allocation of female individuals becoming very low in comparison to the male counterparts.

The present models are not meant to replace existing and very accurate (John Albanese, 2013) models, but to provide alternatives to be used in different contexts, such as the case of commingled, incomplete, and/or fragmented remains, when needed. As stated by Curate et al. (2021), a potential problem with the presented models refers to the chronology of the training sample (late 19th-early 20th centuries), warranting cauting, as is observable by accuracy of allocations in the test sample, in estimating sex in recently deceased individuals.

As a final remark, and as patent in Albanese's (2013) work, the creation of bigger and more generic and representative samples will wield better and more robust results where sex estimation is concerned.

As always statistical models are approximations of reality based on the data used to develop them and caution is always advised with their use.

6. References

Albanese, J. (2010). A Critical Review of the Methodology for the Study of Secular Change Using Skeletal Data. *Ontario Archaeology*, *85-88*, 139–155. https://www.semanticscholar.org/paper/660e377cc1bc35030d4f9b8049e777e6f8b6c63

Albanese, J. (2013). A method for estimating sex using the clavicle, humerus, radius, and ulna. *Journal of Forensic Sciences*, *58*(6), 1413–1419. https://doi.org/10.1111/1556-4029.12188

Algee-Hewitt, B., Coelho, C., Navega, D., & Cunha, E. (2020). Chapter 4.5 - Statistical approaches to ancestry estimation: New and established methods for the quantification of cranial variation for forensic casework. In Z. Obertová, A. Stewart, & C. Cattaneo (Eds.), *Statistics and Probability in Forensic Anthropology* (pp. 227–247). Academic Press. https://doi.org/10.1016/B978-0-12-815764-0.00018-6

Ammer, S., d'Oliveira Coelho, J., & Cunha, E. M. (2019). Outline Shape Analysis on the Trochlear Constriction and Olecranon Fossa of the Humerus: Insights for Sex Estimation and a New Computational Tool. *Journal of Forensic Sciences*, *64*(6), 1788–1795. https://doi.org/10.1111/1556-4029.14096

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

Berner, M., Sládek, V., Holt, B., Niskanen, M., & Ruff, C. B. (2017). Sexual Dimorphism. In *Skeletal Variation and Adaptation in Europeans* (pp. 133–161). John Wiley & Sons, Inc. https://doi.org/10.1002/9781118628430.ch6

Bethard, J. D., & VanSickle, C. (2020). Chapter 3 - Applications of sex estimation in paleoanthropology, bioarchaeology, and forensic anthropology. In A. R. Klales (Ed.), *Sex Estimation of the Human Skeleton* (pp. 25–34). Academic Press. https://doi.org/10.1016/B978-0-12-815767-1.00003-1

Betti, L. (2017). Human Variation in Pelvic Shape and the Effects of Climate and Past Population History. *Anatomical Record* , *300*(4), 687–697. https://doi.org/10.1002/ar.23542

Bidmos, M. A., & Mazengenya, P. (2021). Accuracies of discriminant function equations for sex estimation using long bones of upper extremities. *International Journal of Legal Medicine*, *135*(3), 1095–1102. https://doi.org/10.1007/s00414-020-02458-y

Black, S., Randolph-Quinney, P., & Mallett, X. (2009). Anthropology. In *Wiley Encyclopedia of Forensic Science* (pp. 1–28). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470061589.fsa058.pub2

Bruner, E., Manzi, G., & Arsuaga, J. L. (2003). Encephalization and allometric trajectories in the genus Homo: evidence from the Neandertal and modern lineages. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(26), 15335–15340. https://doi.org/10.1073/pnas.2536671100

Byrd, J. E., & Adams, B. J. (2003). Osteometric sorting of commingled human remains. *Journal of Forensic Sciences*, *48*(4), 717–724. https://doi.org/10.1520/jfs2002189

Christensen, A. M., Passalacqua, N. V., & Bartelink, E. J. (2019). Chapter 9 - Ancestry estimation. In A. M. Christensen, N. V. Passalacqua, & E. J. Bartelink (Eds.), *Forensic Anthropology (Second Edition)* (pp. 271–306). Academic Press. https://doi.org/10.1016/B978-0-12-815734-3.00009-9

Cowal, L. S., & Pastor, R. F. (2008). Dimensional variation in the proximal ulna: evaluation of a metric method for sex assessment. *American Journal of Physical Anthropology*, *135*(4), 469–478. https://doi.org/10.1002/ajpa.20771

Cunha, E., & Wasterlain, S. (2007). The Coimbra identified osteological collections. In G. Grupe & J. Peters (Eds.), *Skeletal series and their socio economic context* (pp. 23–33). Rahden/Westf.

Curate, F., Coelho, J., Gonçalves, D., Coelho, C., Ferreira, M. T., Navega, D., & Cunha, E. (2016). A method for sex estimation using the proximal femur. *Forensic Science International*, *266*, 579.e1–e579.e7. https://doi.org/10.1016/j.forsciint.2016.06.011

Curate, F., Mestre, F., & Garcia, S. J. (2021). Sex assessment with the radius in Portuguese skeletal populations (late 19th - early to mid 20th centuries). *Legal Medicine* , *48*, 101790. https://doi.org/10.1016/j.legalmed.2020.101790

Curate, F., Umbelino, C., Perinha, A., Nogueira, C., Silva, A. M., & Cunha, E. (2017). Sex determination from the femur in Portuguese populations with classical and machine-learning classifiers. *Journal of Forensic and Legal Medicine*, *52*, 75–81. https://doi.org/10.1016/j.jflm.2017.08.011

Cuzzullin, M. C., Curate, F., Freire, A. R., Costa, S. T., Prado, F. B., Daruge Junior, E., Cunha, E., & Rossi, A. C. (2020). Validation of anthropological measures of the human femur for sex estimation in Brazilians. *The Australian Journal of Forensic Sciences*, 1–14. https://doi.org/10.1080/00450618.2020.1729411

DeSilva, R., Flavel, A., & Franklin, D. (2014). Estimation of sex from the metric assessment of digital hand radiographs in a Western Australian population. *Forensic Science International*, *244*, 314.e1–e7. https://doi.org/10.1016/j.forsciint.2014.08.019

d'Oliveira Coelho, J., & Curate, F. (2019). CADOES: An interactive machine-learning approach for sex estimation with the pelvis. *Forensic Science International*, *302*, 109873. https://doi.org/10.1016/j.forsciint.2019.109873

d'Oliveira Coelho, J., Curate, F., & Navega, D. (2020). Chapter 5.1 - Osteomics: Decision support systems for forensic anthropologists. In Z. Obertová, A. Stewart, & C. Cattaneo (Eds.), *Statistics and Probability in Forensic Anthropology* (pp. 259–273). Academic Press. https://doi.org/10.1016/B978-0-12-815764-0.00005-8

El Morsi, D. A., & Al Hawary, A. A. (2013). Sex determination by the length of metacarpals and phalanges: X-ray study on Egyptian population. *Journal of Forensic and Legal Medicine*, *20*(1), 6–13. https://doi.org/10.1016/j.jflm.2012.04.020

Falsetti, A. B. (1995). Sex assessment from metacarpals of the human hand. *Journal of Forensic Sciences*, *40*(5), 774–776. https://www.ncbi.nlm.nih.gov/pubmed/7595320

Ferreira, M. T., Coelho, C., Makhoul, C., Navega, D., Gonçalves, D., Cunha, E., & Curate, F. (2021). New data about the 21st Century Identified Skeletal Collection (University of Coimbra, Portugal). *International Journal of Legal Medicine*, *135*(3), 1087–1094. https://doi.org/10.1007/s00414-020-02399-6

Ferreira, M. T., Vicente, R., Navega, D., Gonçalves, D., Curate, F., & Cunha, E. (2014). A new forensic collection housed at the University of Coimbra, Portugal: The 21st century identified skeletal collection. *Forensic Science International*, *245*, 202.e1–e5. https://doi.org/10.1016/j.forsciint.2014.09.021

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*(2), 365–372. https://doi.org/10.1007/s00414-014-1083-0

Gillet, C., Costa-Mendes, L., Rérolle, C., Telmon, N., Maret, D., & Savall, F. (2020). Sex estimation in the cranium and mandible: a multislice computed tomography (MSCT) study using anthropometric and geometric morphometry methods. *International Journal of Legal Medicine*, *134*(2), 823–832. https://doi.org/10.1007/s00414-019-02203-0

Glennie, R. (2021). Reproducible Research with R and RStudio (3rd Edition) by Christopher Gandrud. *Journal of Agricultural, Biological, and Environmental Statistics*, *26*(1), 128–129. https://doi.org/10.1007/s13253-020-00418-y

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.

Gunz, P., Neubauer, S., Maureille, B., & Hublin, J.-J. (2010). Brain development after birth differs between Neanderthals and modern humans. *Current Biology: CB*, *20*(21), R921–R922. https://doi.org/10.1016/j.cub.2010.10.018

Klales, A. R. (2020). Chapter 16 - MorphoPASSE: Morphological pelvis and skull sex estimation program. In A. R. Klales (Ed.), *Sex Estimation of the Human Skeleton* (pp.

271–278). Academic Press. https://doi.org/10.1016/B978-0-12-815767-1.00016-X

Klales, A. R., Ousley, S. D., & Passalacqua, N. V. (2020). Chapter 13 - Statistical approaches to sex estimation. In A. R. Klales (Ed.), *Sex Estimation of the Human Skeleton* (pp. 203–217). Academic Press. https://doi.org/10.1016/B978-0-12-815767-1.00013-4

Kranioti, E. F., & Michalodimitrakis, M. (2009). Sexual dimorphism of the humerus in contemporary Cretans--a population-specific study and a review of the literature*. *Journal of Forensic Sciences*, *54*(5), 996–1000. https://doi.org/10.1111/j.1556-4029.2009.01103.x

Kuhn, M. (2008). Building Predictive Models inRUsing thecaretPackage. *Journal of Statistical Software*, *28*(5). https://doi.org/10.18637/jss.v028.i05

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer, New York, NY. https://doi.org/10.1007/978-1-4614-6849-3

Langley, N. R., Jantz, L. M., Ousley, S. D., Jantz, R. L., & Milner, G. (2016). Data collection procedures for forensic skeletal material 2.0. *University of Tennessee and Lincoln Memorial University*. http://www.academia.edu/download/45990086/DCP20_webversion.pdf

Langley, N. R., & Jantz, R. L. (2020). Secular change. In A. R. Klales (Ed.), *Sex Estimation of the Human Skeleton* (pp. 295–306). Academic Press. https://doi.org/10.1016/B978-0-12-815767-1.00018-3

Larose, D. T. (2015). *Data Mining and Predictive Analytics*. John Wiley & Sons. https://play.google.com/store/books/details?id=0HJuBwAAQBAJ

Lee, U.-Y., Kim, I.-B., & Kwak, D.-S. (2015). Sex determination using discriminant analysis of upper and lower extremity bones: New approach using the volume and surface area of digital model. *Forensic Science International*, *253*, 135.e1–e4. https://doi.org/10.1016/j.forsciint.2015.05.017

Márquez-Grant, N. (2015). An overview of age estimation in forensic anthropology: perspectives and practical considerations. *Annals of Human Biology*, *42*(4), 308–322. https://doi.org/10.3109/03014460.2015.1048288

Martin, R. (1928). *Lehrbuch der Anthropologie in Systematischer Darstellung mit Besonderer Berücksichtigung der anthropologischen Methoden für Studierende, Artze und Forschungsreisende, Kraniologie, Osteologie* (Vol. 2). Gustav Fischer.

Messer, D. L., & Getz, S. M. (2020). Chapter 5 - Effect of sex misclassification on the skeletal biological profile. In A. R. Klales (Ed.), *Sex Estimation of the Human Skeleton* (pp. 53–72). Academic Press. https://doi.org/10.1016/B978-0-12-815767-1.00005-5

Navega, D., Coelho, C., Vicente, R., Ferreira, M. T., Wasterlain, S., & Cunha, E. (2015). AncesTrees: ancestry estimation with randomized decision trees. *International Journal of Legal Medicine*, *129*(5), 1145–1153. https://doi.org/10.1007/s00414-014-1050-9

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

Neto, M. A. M. (1957). Estudo osteométrico do antebraço nos portugueses 1- rádio. *Contribuições para o estudo da Antropologia Portuguesa*, *6*(6), 143–222. http://hdl.handle.net/10316.2/35192

Nikita, E., & Nikitas, P. (2020). Sex estimation: a comparison of techniques based on binary logistic, probit and cumulative probit regression, linear and quadratic discriminant analysis, neural networks, and naïve Bayes classification using ordinal variables. *International Journal of Legal Medicine*, *134*(3), 1213–1225. https://doi.org/10.1007/s00414-019-02148-4

Nikitovic, D. (2018). Sexual dimorphism (humans). In *The International Encyclopedia of Biological Anthropology* (pp. 1–4). John Wiley & Sons, Inc. https://doi.org/10.1002/9781118584538.ieba0443

Peckmann, T. R., Logar, C., Garrido-Varas, C. E., Meek, S., & Pinto, X. T. (2016). Sex determination using the mesio-distal dimension of permanent maxillary incisors and canines in a modern Chilean population. *Science & Justice: Journal of the Forensic Science Society*, *56*(2), 84–89. https://doi.org/10.1016/j.scijus.2015.10.002

Perini, T. A., Oliveira, G. L. de, Ornellas, J. dos S., & Oliveira, F. P. de. (2005). Cálculo do erro técnico de medição em antropometria. *Revista Brasileira de Medicina do Esporte*, *11*(1), 81–85. https://doi.org/10.1590/S1517-86922005000100009

Petaros, A., Caplova, Z., Verna, E., Adalian, P., Baccino, E., de Boer, H. H., Cunha, E., Ekizoglu, O., Ferreira, M. T., Fracasso, T., Kranioti, E. F., Lefevre, P., Lynnerup, N., Ross, A., Steyn, M., Obertova, Z., & Cattaneo, C. (2021). Technical Note: The Forensic Anthropology Society of Europe (FASE) Map of Identified Osteological Collections. *Forensic Science International*, *328*, 110995. https://doi.org/10.1016/j.forsciint.2021.110995

Pinheiro, J., & Cunha, E. (2006). Forensic Investigation of Corpses in Various States of Decomposition. In A. Schmitt, E. Cunha, & J. Pinheiro (Eds.), *Forensic Anthropology* *and Medicine: Complementary Sciences From Recovery to Cause of Death* (pp. 159–195). Humana Press. https://doi.org/10.1007/978-1-59745-099-7_7

R Core Team, R., & Others. (2013). *R: A language and environment for statistical computing*. R foundation for statistical computing Vienna, Austria.

Reddy, B. B., & Doshi, M. A. (2017). Sex determination from adult human humerus by discriminant function analysis. *International Journal of Research in Medical Sciences*, *5*(9), 3891–3897. https://doi.org/10.18203/2320-6012.ijrms20173702

Rogers, N. L., Flournoy, L. E., & McCormick, W. F. (2000). The rhomboid fossa of the clavicle as a sex and age estimator. *Journal of Forensic Sciences*, *45*(1), 61–67. https://doi.org/10.1520/JFS14641J

Rosenberg, K. R., & DeSilva, J. M. (2017). Evolution of the Human Pelvis. *Anatomical Record* , *300*(5), 789–797. https://doi.org/10.1002/ar.23580

Ross Quinlan, J. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann. https://play.google.com/store/books/details?id=HExncpjbYroC

RStudio Team. (2020). *RStudio: Integrated Development Environment for R*. RStudio, PBC. http://www.rstudio.com/

Ruff, C. (1987). Sexual dimorphism in human lower limb bone structure: relationship to subsistence strategy and sexual division of labor. *Journal of Human Evolution*, *16*(5), 391–416. https://doi.org/10.1016/0047-2484(87)90069-8

Santos, F., Guyomarc'h, P., Cunha, E., & Brůžek, J. (2020). Chapter 15 - DSP: A probabilistic approach to sex estimation free from population specificity using innominate measurements. In A. R. Klales (Ed.), *Sex Estimation of the Human Skeleton* (pp. 243–269). Academic Press. https://doi.org/10.1016/B978-0-12-815767-1.00015-8

Silva, A. M. (1995). Sex assessment using the calcaneus and talus. *Antropologia Portuguesa*, *13*, 107–119. URI:http://hdl.handle.net/10316.2/30547

Skiena, S. S. (2017). Mathematical Models. In S. S. Skiena (Ed.), *The Data Science Design Manual* (pp. 201–236). Springer International Publishing. https://doi.org/10.1007/978-3-319-55444-0_7

Spradley, M. K. (2016). Metric Methods for the Biological Profile in Forensic Anthropology: Sex, Ancestry, and Stature. *Academic Forensic Pathology*, *6*(3), 391–399. https://doi.org/10.23907/2016.040

Spradley, M. K., & Jantz, R. L. (2011). Sex estimation in forensic anthropology: skull versus postcranial elements. *Journal of Forensic Sciences*, *56*(2), 289–296. https://doi.org/10.1111/j.1556-4029.2010.01635.x

Ubelaker, D. H. (2014). Osteology Reference Collections. In C. Smith (Ed.), *Encyclopedia of Global Archaeology* (pp. 5632–5641). Springer New York. https://doi.org/10.1007/978-1-4419-0465-2_159

Ubelaker, D. H., & DeGaglia, C. M. (2017). Population variation in skeletal sexual dimorphism. *Forensic Science International*, *278*, 407.e1–e407.e7. https://doi.org/10.1016/j.forsciint.2017.06.012

Ubelaker, D. H., & DeGaglia, C. M. (2020). Chapter 17 - Factors of population variation in sex estimation methodology. In A. R. Klales (Ed.), *Sex Estimation of the Human*

Skeleton (pp. 281–293). Academic Press. https://doi.org/10.1016/B978-0-12-815767-1.00017-1

Wasterlain, S. (2000). *Morphé : análise das proporções entre os membros, dimorfismo sexual e estatura de uma amostra da colecção de esqueletos identificados do Museu Antropológico da Universidade de Coimbra* [Masters dissertation in Human Evolution, University of Coimbra]. https://eg.uc.pt/handle/10316/9870

Wells, J. C. K. (2012). Sexual dimorphism in body composition across human populations: associations with climate and proxies for short- and long-term energy supply. *American Journal of Human Biology: The Official Journal of the Human Biology Council*, *24*(4), 411–419. https://doi.org/10.1002/ajhb.22223

Zeman, T., & Beňuš, R. (2020). Chapter 4.6 - Stature estimation. In Z. Obertová, A. Stewart, & C. Cattaneo (Eds.), *Statistics and Probability in Forensic Anthropology* (pp. 249–256). Academic Press. https://doi.org/10.1016/B978-0-12-815764-0.00020-4

Zuckerman, M. K., & Crandall, J. (2019). Reconsidering sex and gender in relation to health and disease in bioarchaeology. *Journal of Anthropological Archaeology*, *54*, 161–171. https://doi.org/10.1016/j.jaa.2019.04.001

Langley, N. R., Jantz, L. M., Ousley, S. D., Jantz, R. L., & Milner, G. (2016). *Data collection procedures for forensic skeletal material 2.0*. University of Tennessee and Lincoln Memorial.

7. Annexes

Table A1 - Data collected on the first measurement to calculate intra-observer error

id	hml	hmc	hvdh	hhdh	hcb	heb	uml	upl	umc	ubsn	uob	udaw
114	300.0	61.5	43.3	40.3	42.7	59.2	244.0	217.0	33.0	24.2	23.2	16.5
177	278.0	46.0	33.4	30.4	33.3	47.2	214.0	189.0	27.0	15.9	18.9	13.9
179	322.5	62.0	46.3	42.5	44.4	60.8	247.5	219.0	35.0	24.1	24.8	18.3
180	285.0	55.5	38.4	33.5	38.4	54.2	235.5	211.0	35.0	20.8	19.4	15.7
182	290.5	55.0	40.1	36.1	40.1	54.8	231.5	205.0	32.5	20.9	22.7	15.4
183	317.5	70.0	45.1	40.8	46.0	63.1	263.5	230.5	42.0	25.3	26.5	19.3
184	309.5	65.5	47.0	43.4	44.7	59.7	246.5	217.0	40.0	25.7	25.1	16.4
190	310.5	54.5	41.8	38.4	39.2	50.6	238.5	213.0	32.0	19.2	21.0	14.0
191	292.5	61.0	39.6	37.7	41.7	59.3	247.5	221.0	32.0	22.8	22.9	14.4
192	381.5	60.0	39.9	39.4	42.1	56.3	233.0	205.0	35.0	22.8	25.9	17.0
199	333.0	64.0	47.2	43.1	44.9	59.3	257.0	228.0	43.0	24.9	25.3	15.7
202	303.5	61.0	39.7	NA	38.3	54.0	246.5	218.0	35.0	21.1	23.5	14.8
203	318.5	58.0	42.3	38.3	41.7	58.3	241.0	215.0	34.0	23.5	22.8	13.5
204	326.0	61.0	49.0	42.5	46.5	63.6	263.5	229.5	35.5	24.1	26.1	19.1
205	329.5	57.0	42.3	NA	41.3	58.6	272.0	237.5	37.5	22.4	26.3	15.5
206	308.5	59.0	39.2	36.4	38.4	54.0	236.0	211.0	30.0	20.1	23.1	14.2
207	295.5	52.5	40.7	36.5	38.7	55.5	233.5	209.0	31.0	20.4	21.1	16.0
208	278.0	64.0	41.6	40.4	41.6	56.9	236.1	208.0	35.0	21.6	24.0	14.9
209	298.0	52.0	39.6	37.1	38.2	51.8	226.0	203.0	30.0	20.8	21.8	15.7
211	291.0	54.0	38.5	35.7	39.0	54.5	241.5	216.0	32.0	19.8	21.1	15.1
213	321.5	68.5	47.5	45.3	47.7	65.0	263.0	233.0	37.0	28.0	26.7	18.9
215	293.0	52.0	37.0	35.5	39.5	52.6	217.5	192.0	31.0	22.4	23.2	16.0
217	330.5	67.0	46.1	43.2	46.1	61.8	264.0	237.5	38.0	25.0	27.1	15.3
222	310.5	53.0	39.8	37.6	37.7	54.5	230.5	204.5	31.0	21.5	21.2	11.8
225	284.0	51.0	38.1	35.4	35.9	51.1	225.0	197.0	30.5	18.8	20.8	11.6
229	300.0	56.0	40.2	38.1	39.4	53.2	226.5	201.5	32.5	22.5	21.8	14.0
231	303.5	71.0	48.6	43.1	47.6	66.9	258.5	230.0	33.0	26.6	25.9	16.6
235	303.5	66.0	43.8	42.1	42.9	65.0	238.5	206.0	37.0	23.2	24.7	15.7
236	265.5	56.0	37.8	34.5	38.3	46.9	213.0	189.5	29.0	19.3	19.4	15.5
237	309.5	67.5	46.4	41.9	44.4	62.6	240.0	209.5	33.5	23.7	23.6	15.0

Table A2 - Data collected on the second measurement to calculate intra-observer error

id	hml	hmc	hvdh	hhdh	hcb	heb	uml	upl	umc	ubsn	uob	udaw

114	300.5	61.5	43.3	41.1	42.6	59.2	244.5	216.0	33.0	24.3	23.3	16.5
177	278.0	46.0	33.4	30.9	33.7	47.2	214.5	189.0	26.5	16.4	18.8	13.9
179	323.0	62.0	46.5	42.8	43.7	60.8	247.5	219.0	35.0	23.7	24.9	18.4
180	285.5	55.5	38.4	33.6	38.6	54.2	236.0	211.0	34.5	20.6	20.5	15.9
182	291.0	55.3	39.9	36.3	40.3	54.8	232.0	205.0	32.5	21.2	23.2	15.5
183	318.0	70.0	45.0	41.8	45.6	63.1	264.5	231.5	42.0	25.4	27.0	19.5
184	310.0	65.5	46.9	43.3	44.4	59.7	247.0	217.5	40.0	26.0	25.3	16.4
190	311.0	54.5	42.0	38.4	39.2	50.4	239.0	213.5	32.0	20.6	21.1	13.9
191	273.0	61.0	39.7	37.8	41.9	59.9	247.5	221.0	32.0	23.3	22.9	13.8
192	381.5	59.5	40.1	39.4	42.4	56.3	233.5	205.0	35.0	22.8	25.9	17.0
199	333.5	63.5	47.1	43.4	45.1	59.3	257.0	228.0	42.0	25.0	24.7	15.2
202	303.5	60.5	39.5	NA	39.2	53.9	247.0	219.0	34.5	21.5	23.4	15.0
203	319.0	58.0	42.4	38.1	42.5	58.3	241.0	214.0	33.5	23.4	22.8	13.4
204	326.0	61.0	49.2	42.3	46.4	64.1	264.0	230.0	35.5	24.4	27.0	19.1
205	330.0	56.5	42.6	NA	41.6	59.0	272.5	238.0	37.0	23.0	26.2	15.3
206	308.5	58.0	38.9	36.5	37.9	54.0	236.0	211.5	30.0	20.2	23.5	14.3
207	296.0	52.5	40.7	36.9	38.5	55.5	234.0	209.0	30.5	20.5	21.0	16.2
208	279.0	63.0	41.6	40.1	42.1	57.0	236.5	208.0	34.0	21.7	22.4	15.1
209	297.5	51.5	39.4	36.9	38.3	51.9	226.0	202.0	29.0	20.8	21.8	15.8
211	291.5	53.5	38.4	35.9	39.9	54.9	242.0	216.0	31.5	19.9	20.9	15.1
213	320.5	68.5	47.5	45.1	47.9	65.1	263.5	233.0	37.0	27.6	26.4	19.1
215	293.5	52.0	37.3	35.5	39.5	52.5	217.5	191.0	31.0	22.7	23.3	15.9
217	331.0	66.5	46.2	43.4	46.6	61.8	265.5	238.0	37.5	25.0	27.0	15.4
222	310.5	53.0	39.8	37.1	38.0	54.7	231.0	204.0	31.0	21.2	21.2	11.8
225	284.0	50.5	38.7	35.6	35.8	50.9	225.0	196.0	30.0	18.9	20.8	11.4
229	301.0	56.0	40.3	38.1	39.1	53.3	226.5	202.0	32.0	22.5	21.8	14.0
231	304.0	70.5	48.6	43.4	47.8	66.9	258.5	231.0	33.0	26.6	26.2	16.6
235	303.5	65.5	44.0	42.7	42.6	64.9	238.5	207.0	36.0	23.3	24.6	15.9
236	266.5	55.5	37.8	35.1	39.1	46.6	214.0	190.0	29.0	19.3	19.4	15.5
237	309.5	67.5	46.8	41.9	44.7	62.8	240.0	209.5	33.5	24.2	25.1	14.9

Table A3 - Data from Neto (1957) to calculate inter-observer error

id	rml	rmcdd	rhc	rnc	rdew

114	225.0	39.0	66.0	45.0	32.0
177	198.5	28.5	53.0	33.5	25.5
179	233.0	40.0	67.0	45.0	31.5
180	218.0	35.0	56.0	37.5	31.0
182	212.0	33.0	62.0	38.5	29.0
183	242.0	44.0	73.0	52.0	34.0
184	230.0	46.0	68.0	50.0	33.0
190	222.0	35.0	64.0	41.5	30.0
191	230.0	38.0	65.0	39.0	32.0
192	217.5	35.0	68.0	45.0	31.0
199	237.5	41.0	68.0	47.0	33.0
202	228.0	37.0	56.0	39.0	31.0
203	225.0	39.0	66.0	40.0	32.0
204	244.5	42.0	71.0	45.0	38.0
205	251.0	36.0	67.0	45.5	29.0
206	218.0	36.0	59.0	44.0	28.5
207	216.0	32.0	60.0	38.5	30.0
208	255.0	40.0	66.0	46.0	30.0
209	213.0	34.0	60.0	38.0	30.5
211	222.0	33.0	64.0	42.5	28.0
213	244.0	42.0	75.0	47.0	36.0
215	197.5	35.0	61.0	42.0	30.0
217	246.0	42.0	74.5	48.5	34.0
222	212.0	33.0	58.0	40.0	28.0
225	209.0	32.0	55.0	39.0	27.0
229	212.0	35.5	64.0	41.5	30.0
231	239.0	42.0	71.0	47.5	34.0
235	222.0	45.0	66.0	51.0	33.0
236	192.0	30.0	56.0	36.5	27.5
237	223.0	43.0	68.0	45.0	34.0

Table A4 - Data collected to calculate inter-observer error (CISC)

id	rml	rmcdd	rhc	rnc	rdew

114	224.0	39.0	65.0	45.0	32.1
177	198.5	28.5	52.5	33.5	25.0
179	233.0	40.0	67.0	45.0	31.1
180	217.5	35.5	55.0	37.5	30.4
182	212.5	33.0	62.0	38.5	29.1
183	242.0	44.0	73.0	52.0	33.9
184	230.0	46.0	65.0	50.0	33.1
190	221.5	35.0	64.0	42.0	29.7
191	230.0	38.5	64.0	39.0	31.8
192	217.5	35.5	61.0	45.0	31.2
199	237.5	41.5	69.0	47.0	32.2
202	228.0	37.0	56.0	39.0	29.8
203	225.0	39.0	65.0	40.0	32.1
204	244.5	42.0	69.0	45.0	37.1
205	251.5	36.0	69.0	46.0	28.5
206	218.0	36.5	61.0	44.0	27.8
207	216.0	32.0	60.0	38.5	29.2
208	254.0	40.5	67.0	46.0	29.9
209	213.5	34.0	61.0	38.0	30.3
211	223.0	33.0	63.0	42.5	27.6
213	244.5	42.0	75.0	47.0	35.5
215	197.5	35.0	61.0	42.0	29.7
217	246.5	42.5	75.0	48.5	33.6
222	212.5	33.0	58.0	40.0	28.1
225	209.5	32.0	56.0	39.0	26.5
229	217.5	35.5	64.0	41.5	30.0
231	239.0	42.0	68.0	47.5	33.5
235	223.0	44.5	66.0	51.0	32.5
236	192.5	30.0	57.0	36.5	28.0
237	224.0	43.0	68.0	45.0	33.3

Table A5 - Summarized accuracy values for sectioning points

	Humerus	Ulna	Radius
Accuracy CV	79.3 - 94.4	75.4 - 87.1	80.5 - 90.5
Accuracy Test	68.6 - 77.1	60.5 - 79.0	70.6 - 82.4
Females	57.9 - 73.7	25.0 - 85.0	60.0 - 80.0
Males	81.3 - 93.8	66.7 - 100	85.7 - 92.9

Legend: CV - cross-validation; Test - test sample; Female - % of correctly classified females; Male - % of correctly classified males

Table A6 - Summarized accuracy values for LR multivariate models for individual bones

	Humerus(1)	Ulna(3)	Radius(3)
Accuracy CV	95.2	89.4 - 90.2	92.2 - 92.6
Accuracy Test	74.3	79.0 - 81.6	85.3
Females	57.9	65.0 - 70.0	80.0
Males	93.8	94.4	92.9

Legend: CV - cross-validation; Test - test sample; Female - % of correctly classified females; Male - % of correctly classified males

Table A7 - Summarized accuracy values for LR multivariate models for combinations bones

	HU(1)	HR(3)	UR(2)	HUR(1)
Accuracy CV	96.7	95.1 - 96.2	92.9	95.6
Accuracy Test	81.8	81.8	81.8 - 86.4	81.8
Females	69.2	69.2	69.2 - 76.9	69.2
Males	100	100	100	100

Legend: CV - cross-validation; Test - test sample; Female - % of correctly classified females; Male - % of correctly classified males

Table A8 - Summarized accuracy values for SVM multivariate models for individual bones

	Humerus(1)	Ulna(3)	Radius(3)
Accuracy CV	96.1	90.2 - 91.4	92.2 - 93.1
Accuracy Test	77.1	79.0 - 84.2	85.3
Females	63.2	65.0 - 75.0	80.0
Males	93.8	94.4	92.9

Legend: CV - cross-validation; Test - test sample; Female - % of correctly classified females; Male - % of correctly classified males

Table A9 - Summarized accuracy values for SVM multivariate models for combinations bones

	HU(1)	HR(3)	UR(2)	HUR(1)
Accuracy CV	96.2	95.6 - 96.2	92.9	96.2
Accuracy Test	81.8	81.8	81.8 - 86.4	81.8
Females	69.2	69.2	69.2 - 76.9	69.2
Males	100	100	100	100

Legend: CV - cross-validation; Test - test sample; Female - % of correctly classified females; Male - % of correctly classified males