2022

Incorporating metrics and nonmetrics in the development of a population-inclusive sex estimation model using volume rendered CT images of the skull

BOSTON UNIVERSITY

SCHOOL OF MEDICINE

Thesis

INCORPORATING METRICS AND NONMETRICS IN THE DEVELOPMENT OF A POPULATION-INCLUSIVE SEX ESTIMATION MODEL USING VOLUME RENDERED CT IMAGES OF THE SKULL

by

SAMANTHA R. KELLEY

B.A., Boston University, 2019

Submitted in partial fulfillment of the

requirements for the degree of

Master of Science

2022

Approved by

First Reader

Sean D. Tallman, Ph.D., RPA
Assistant Professor of Anatomy and Neurobiology

Second Reader

Tara L. Moore, Ph.D.
Associate Professor of Anatomy and Neurobiology

ACKNOWLEDGMENTS

Thank you, a thousand times over, to Dr. Sean Tallman for all his assistance and guidance throughout this entire process. His patience, support, and SPSS know-how were integral to the success of this research. To Dr. Tara Moore and the faculty of the Forensic Anthropology Department at Boston University, thank you for giving me the opportunity to continue my education and the ability and tools I need to chase my dream of becoming a forensic anthropologist. Thank you, also, to the members of my cohort for the late-night study sessions and pep talks, especially Bella Liggett and Alyssa Zamora who provided the moral support and coffee breaks I needed to make it through grad school.

I would also like to acknowledge my mom for her unwavering encouragement and understanding through this process. Thank you for letting me use your office for the last two years and for all of the wisdom you've imparted to me the last twenty-four. I literally would not be here if it weren't for you, but I definitely wouldn't be able to pursue my dream of becoming a forensic anthropologist if it weren't for your steadfast support and love.

INCORPORATING METRICS AND NONMETRICS IN THE DEVELOPMENT OF A POPULATION-INCLUSIVE SEX ESTIMATION MODEL USING VOLUME RENDERED CT IMAGES OF THE SKULL

SAMANTHA R. KELLEY

ABSTRACT

In forensic anthropology, methods for estimating sex from the skeleton have historically been developed on skeletal collections comprised exclusively U.S. Black and white populations and thus inherently rely on ancestry estimation as a foundational component for the construction of the biological profile. However, these population-specific methods of sex estimation fundamentally limit the applicability, accuracy, and reliability of the method for use on remains of diverse population affiliations (ancestry). A reliable and population-inclusive method for estimating sex in the skeleton can serve as a useful tool for forensic investigators, especially in cases with unidentified remains where the population affiliation is indeterminate and in light of recent discourse involving the removal of ancestry estimation from the biological profile as a whole This study assessed the reliability of population-inclusive nonmetric and metric sex estimation from 3D-volume rendered computer tomography (CT) scans of the skull. The study sample was comprised of 431 individuals (242 males and 189 females) from the New Mexico Decedent Imaging Database (NMDID) and included a relatively equal distribution of African American, Asian American, European American, Latin American, and Native American population affinities. The images were obtained from the CT slices using 3D-reconstructions and volume rendering technique (VRT) in the Digital Imaging and

Communications in Medicine (DICOM) viewer, exported to Meshmixer™, and then processed to isolate the skull from the postcranial skeleton and remove identifying objects. In Meshmixer™, nonmetric traits were scored following Buikstra and Ubelaker (1994) and Walker (2008) and included the supraorbital ridge/glabella, supraorbital margin, mastoid process, mental eminence, and nuchal crest. The metric traits, following Spradley and Jantz (2011), included 18 points of measurement of the cranium and mandible. Binary logistic regression (BLR) and discriminant function analyses (DFA) were used to produce models and probabilities from the nonmetric and metric data respectively and an additional binary logistic regression was developed that combined both the nonmetric and metric data. Overall, the population-inclusive nonmetric and metric model produced classification accuracies that ranged from 81-87% and 86.7-87% respectively, and performed as well as population-specific models in estimating sex and were not significantly different from population-specific accuracies. When some of the population-specific models were applied across the sample populations, particularly the European American model, the classification accuracy was significantly reduced relative to the population-inclusive model. Intraobserver error was assessed for the nonmetric and metric data collection and confirmed that the nonmetric and metric methods of data collection for the volume-rendered images were consistent. The results of this study indicate that a population-inclusive nonmetric and metric models of sex estimation using the skull can be used in place of more traditional population-specific models in cases where ancestry is unknown, indeterminate, or in the event ancestry is removed from the biological profile.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

BLR	Binary Logistic Regression
BU	Boston University
CT	Computed Tomography
DFA	Discriminant Function Analysis
DICOM	Digital Imaging and Communications in Medicine
MDCT	Multidirectional Computed Tomography
MSCT	Multislice Computed Tomography
NIJ	National Institute of Justice
NMDID	New Mexico Decedent Image Database
OMI	Office of the Medical Investigator
PVE	Partial Volume Effect
VRT	Volume Rendering Technique

CHAPTER ONE: INTRODUCTION

Assigned skeletal sex estimation is an integral component and crucial step in the development of the biological profile of an individual in bioarchaeological, medico-legal, and other anthropological contexts (Klales 2020, 2013; Klales et al. 2012; Rogers 2005Spradley & Jantz 2011; Stewart 1979; Tallman 2019). In bioarchaeological analyses, sex estimation from the skeletal remains can aid researchers in reconstructing the biology and behavior of past populations (Albanese 2013; Walrath et al. 2004). In forensic anthropology, sex estimation often serves as the foundation on which other assumptions and identifications are made and additionally dictates the methods that are subsequently utilized to estimate age and stature (Klales 2020; Tallman 2019). Estimating binary sex in an unidentified set of remains can also be incredibly valuable because this identification—male or female—helps to narrow down the search parameters by approximately 50% when comparing the case to missing person reports (Spradley 2016). However, this statement is inherently incorrect because sex is not a binary process (Blackless et al. 2000). Moreover, the sex of an individual can often not be estimated with high accuracy until ancestry is estimated (Klales 2020)—that is, sexual dimorphism differs between certain populations; however, this variation is unlikely to fall along ancestral (i.e., race) lines. This is due to the prevailing theory that population variation affects levels of sexual dimorphism present in the skeleton (Klales 2020; Tallman 2019; Ubelaker & DeGalgia 2017), leading researchers to develop population-specific metric and nonmetric methods of sex estimation using the crania (e.g., Acsádi & Nemeskéri 1970; Broca 1875; Buikstra & Ubelaker 1994; Dayal et al. 2008; Garvin et al. 2014;

Graw et al. 1999; Krogman 1955; Krogman & Işcan 1986; Lewis & Garvin 2016; Martin & Knussman 1988; Rogers 2005; Spradley & Jantz 2011; Tallman 2019; Tallman & Go 2018; Walker 2008; Walrath et al. 2004) and post-cranial skeleton (e.g., Buikstra & Ubelaker 1994; Karsten 2017; Klales et al. 2012; Letterman 1941; Patterson & Tallman 2019; Phenice 1969; Rogers 1999, Rogers et al. 2000; Spradley & Jantz 2011; Vance et al. 2011; Walker 2005; Washburn 1948).

As will be discussed in chapter two, research regarding gender diverse individuals in the forensic anthropological record is limited (Kincer & Tallman 2019; Klales 2020; Tallman et al. 2021) and there is a lack of discipline-wide standards on how to report or interpret the relationship between gender identity and biological sex based on skeletal morphology (Klales 2020). Furthermore, the conventional sex estimations currently utilized by forensic anthropologists in case work, generally estimate sex based on binary (i.e., female or male) which is problematic because the typical binary fails to capture the range of variation in humans that goes beyond the binary biological sex descriptors.

The pelvis is generally accepted as the best indicator of assigned sex at birth (Buikstra & Ubelaker 1994; Klales et al. 2012, Krogman & Işcan 1986; Letterman 1941; Phenice 1969; Walker 2005), followed by the long bones (Curate et al. 2016; Rogers 1999; Rogers et al. 2000; Patterson & Tallman 2019; Spradley & Jantz 2011; Vance et al. 2011). And while these postcranial elements provide the highest relative accuracy for estimating sex, rates of differential preservation indicate that they are not always available (Albanese 2013; Waldron 1987). In these cases, the skull (cranium and

mandible) provides the next best indicator of sex (Bass 2005; Krogman & Işcan 1986; Stewart 1979).

The estimation of sex from the skeleton has generally been accomplished through visual observation of the morphology of sex-specific traits or quantitative assessment of known sexually dimorphic metric measurements (Buikstra & Ubelaker 1994; Klales 2020; Ubelaker & DeGalgia 2017), but more recent methods of estimation including discriminate function equations made available with Fordisc (Jantz & Ousely 2013) or molecular examination (Gaballa et al. 2014) are also available. Nonmetric traits, in particular, are important to forensic anthropology because in contexts with fragmentary, incomplete, and poorly preserved remains, the nonmetric data can still be collected (Buikstra & Ubelaker, 1994) and provide a credible method of sex estimation (Rogers 2005). The observable differences between male and female skeletons are often attributed to hormonal action, which is considered one of the greatest mechanisms that affect sex differences (Klales 2020), as well as biomechanical and function constraints with respect to obstetrics, sexual selection, as well as mating preferences (Best et al. 2018; Frayer & Wolpoff 1985). Sexual dimorphism is generally observable in the expression of certain skeletal features and traits which are assigned an ordinal score that indicates a level of gracility (problematically labeled “female-like” or “male-like" if robust) (Acsádi & Nemeskéri 1970; Buikstra & Ubelaker 1994; Tallman 2019). It is important to note that the descriptive terms of “feminine” and “masculine” have historically been used to describe certain morphological expression of skeletal features as seen in the Acsádi and Nemeskéri (1970) and Buikstra and Ubelaker (1994), when in actuality these terms refer

to socially constructed norms that cannot be discerned from the skeleton. The expressions "feminine" and "masculine" signify personality qualities, behaviors, and roles sometimes associated with specific genders and influenced by society and vary by culture (Klales 2020) and are therefore incorrect when applied to the skeleton. Walker (2008) and Klales et al. (2012) (discussed more in chapter two) advocate for and propose a more objective ordinal scoring system from 1 to 5 rather than "feminine" to "masculine", aiming to incorporate a stronger focus on observing the size, shape, robusticity, and gracility of specific traits and the application of statistical probabilities (Klales 2020).

Metric methods of sex estimation are perceived as being more objective than their qualitative counterparts (Christensen et al. 2019; Dirkmaat et al. 2008; Moore 2013). Spradley and Jantz (2011:290) describe sex estimation as the use of "metric traits of the pelvis, skull, or any single bone, or combination of bones" and state explicitly that it "provides an estimate in the form of an error rate or expected classification rate" (Spradley & Jantz 2011:290). Although seemingly superior to nonmetric methods of sex estimation, intra- and inter-observer error rates still exist within metric models and the increased validation and use of regression models with nonmetric methods, means that morphological observation now, too, have classification accuracies, probabilities, and error rates (Klales 2018; Klales et al 2012; Walker 2008).

Regardless of whether or not a nonmetric or metric trait is being utilized, the majority of the skeletal collections with documented histories that are available for researchers to study and develop forensic anthropological identification methods are largely from U.S. Black, white, and Native American population groups (Spradley et al. 2008; Tallman

2019; Tise et al. 2013). The Forensic Anthropology Data Bank (FDB) also has a dearth of data on positively identified Hispanic individuals (or other underrepresented groups) meaning that reliable population-specific sex estimation methods for this specific population are not possible (Tise et al. 2013).

Recently, the role of ancestry estimation in the biological profile has come under scrutiny with some professionals in the field questioning its validity as a component of the biological profile (Adams & Pilloud 2020; Albanese & Saunders 2006; Bethard & Digangi 2020; Blakey 2021; Carson 2006a; DiGangi & Bethard 2021; Edgar 2020; Ross et al. 2020; Spradley & Jantz 2020; Tallman et al. 2021). Given the deficient and absent explanations for the primary mechanisms behind the manifestation of the traits used in ancestry estimation and lack of critical inquiry into their role as tools for reinforcing of the debunked biological race theory, some researchers are calling for an "immediate moratorium on the use of morphoscopic cranial traits in the estimation of ancestry" (Bethard & DiGangi 2020:1791).

Rationale of the Current Study

According to Bethard and DiGangi (2020), the majority of practicing forensic anthropologists operate under the explicit assumption that ancestry is an essential and critical piece of the biological profile, but the methods of ancestry estimation conventionally utilized, especially the morphoscopic methods are problematic and potentially inaccurate. The current study seeks to address the ongoing debate regarding the role of ancestry in the biological profile by proposing a method of sex estimation that does not rely on ancestry estimation. This population-inclusive model would be

applicable in cases where ancestry is unknown, intentionally not estimated in order to mitigate potential for racial biases like the "missing white woman syndrome" (Sommers 2016; Moss 2018), and in light of the debate surrounding the removal of certain ancestry estimation methods from the construction of the biological profile.

Sex estimation should be treated as a separate component and not reliant on ancestry estimation in cases where ancestry is either unknown or intentionally not estimated, in cases where population-specific methods are not available, or in the event that ancestry is removed from the biological profile. Estimation of population affinity (i.e., "ancestry." "race," "bioaffinity") has been made compulsory for human identification, and the development of the biological often depends on population-specific models (Bethard & DiGangi 2020). While some postcranial research has indicated that highly accurate population-inclusive sex estimation (Albanese 2013, 2008) and other identification methods (Albanese 2016) can be developed, population-specific methods continue to be used in most cases, even if the appropriate population-specific method does not exist (Tallman 2019).

Current research may indicate that some level of population variation exists within the skeleton and has the potential to affect expression of sexually dimorphic traits (Klales 2020; Phenice 1969; Spradley & Jantz 2011; Tallman & Go 2018; Walker 2008), but the majority of the methods available to researchers to be employed in the professional sector were developed on U.S. Black and European (U.S. white and English) populations (Klales 2020; Tallman 2019). These population-specific methods are exclusionary in nature and are still being applied across many population affinities or to unidentified

skeletal remains. Until the mechanisms behind population variation in skeletal morphology are better understood or until a comprehensive tool kit of population-specific methods that include models for every possible population (especially in the FDB) exists, population-inclusive models for estimating sex should be developed, validated, and become part of standard practices.

This study sought to develop a population-inclusive model for estimation sex from nonmetric and metric traits of the skull (cranium and mandible). Full-body computed tomography (CT) scans of individuals from the New Mexico Decedent Image Database (NMDID) were compiled into 3D-volume rendered images from which morphoscopic traits could be observed and metric measurements collected. While population-inclusive sex estimation models for the postcrania have been shown to perform with statistical accuracy (Albanese et al. 2016,2008), there is little research with the skull.

Computed Tomography (CT)

In 1979, the Nobel Prize was awarded to Sir Godfrey Hounsfield and Allan McLeod Cormack, who were credited for developing similar medical imaging techniques that led to the modern computed tomography (CT) scanners (Beckmann 2006; Petrik et al. 2006). As technology advanced, the quality and imaging capabilities of the CT technology improved markedly and today, high-resolution images of even the smallest structures in the human body, like the vasculature and tracheobronchial tree, can be imaged and analyzed (Benya 2008).

The National Institute of Biomedical Imaging and Bioengineering (2021) defines CT as a “computerized x-ray imaging procedure in which a narrow beam of x-rays is aimed

at a patient and quickly rotated around the body, producing signals that are processed by the machine's computer to generate cross-sectional images—or "slices"—of the bodies." The slices produced from the x-ray signals are called *tomographic images*, which contain a more detailed picture of whatever is being scanned than a normal x-ray. When successive CT slices are generated, the computer is able to digitally combine, or "stack" them, in order to construct a three-dimensional image of whatever structure has been imaged (e.g., a human skull). This "stacking" process allows researchers to more easily examine structures or identify possibly abnormalities (NIBIB 2021).

The most common form of CT used is multislice CT (MSCT) or multidirectional CT (MDCT), which may capture between 4 and 640 x-ray slices from a variety of angles (Klales 2020:356). The MDCT modality is not only the usual form in clinical settings, but is gaining increasing popularity in medico-legal institutions (Oesterhelwege & Thali 2008).

In forensic anthropology and archaeology, the use of CT scans in forensic anthropology and skeletal research has proved to be a suitable alternative method for collecting data, especially when traditional analysis of dry bone is not possible (Herrera and Tallman 2019; Hughes et al. 2013; Richard et al. 2014; Ross 2004; Stull et al. 2014). In addition to having applications in disaster victim identification (Blau et al. 2008; Brough et al. 2015) and archaeological context (Gardner et al. 2004), this research will expand upon the utility of CT data collection for developing the component of sex estimation in the biological profile.

New Mexico Decedent Image Database (NMDID)

In 2010, the National Institute of Justice (NIJ) awarded the New Mexico Office of the Medical Investigator (OMI) a research grant to determine the supplemental value of CT scans and their potential utility in supplanting traditional autopsies (Berry 2014; Daneshvari Berry & Edgar 2017). The OMI is a centralized medical examiner's office that serves the whole state of New Mexico, and from mid-2010 and mid-2017, any decedent who was routed to the OMI and underwent subsequent autopsy also received a high resolution, full-body CT scan. Although the original grant focused on deaths related to gunshot wounds (GSW), sharp force trauma, and infant deaths, in 2010 the OMI began to routinely scan decedents in addition to the traditional examination. This included any individual who died in a sudden, untimely, or unexpected manner as well as any person found dead and the cause of death was unknown (Daneshvari Berry & Edgar 2017, 2021).

Because of this intensive effort, the NMDID is able to provide de-identified full-body CT scans and corresponding demographic, health lifestyle, and death circumstance data to qualified researchers through an online platform. As of 2021, the database contains data approximately 15,243 decedents with full body scout images equivalent to whole body radiographs and potentially as many as 69 metadata fields (Figure 1.1). Each decedent is associated with two sets of CT scans, augmented for both soft tissue and bone analysis. Each image is comprised of 4,000 axial image slices, with a 512x512 matrix and a slice thickness of 1mm, with a 0.5 mm overlap (Daneshvari Berry & Edgar 2021).

Berry (2014) determined the corresponding metadata fields for the CT images using a virtual Delphi method. Researchers from anthropology, medicine, forensics, informatics, epidemiology, biomedical research, and dentistry were queried and these experts determined 59 variables that were considered “essential” to associated with the CT scans. Berry (2014) added an additional 10 variables to account for demographic specific questions in New Mexico (e.g., marital status, ethnicity, Hispanic subgroup identification, parents, and grandparent’s nativity). In 2016, the NIJ awarded a grant specifically to develop the CT database with the corresponding metadata (Daneshvari Berry & Edgar 2021).

Sixty-nine metadata fields available in NMDID		
Census	**Health**	**Circumstances of Death**
Sex	Cancer	Primary cause of death
Birth year	Congenital and genetic disorders	Contributing cause of death
Death year	Chromosomal abnormalities	Manner of death
Zip code	Dental health as an adult	Identification method
Age in years	Dental health as a child	Bone density
Age in months	Scoliosis	Death city
Gender	History of broken bones	Death county/tribal land
Number of pregnancies	Facial trauma	Cadaver condition
Number of life births	Plastic surgery	Cadaver length
Marital status	Implants	Cadaver weight
Living weight	Radiation therapy	CT scan settings
Living height	Medical diagnoses	Time delay after death
Race	Notes on Medical diagnoses	Name of person entering data
Tribal affiliation	Medications	
Ethnicity	Substance usage	
Hispanic identification	Quit date of substance	
Birth weight	Years of substance usage	
Birth weight category	Tobacco type	
Birthplace	Tobacco usage	
Years in the US	Drinking status	
Mother's birthplace	Dietary pattern	
Father's birthplace	Occupations	
Mother's mother birthplace	Duration of occupations	
Mother's father birthplace	Activities	
Father's mother birthplace	Strenuous lifting	
Father's father birthplace	Education level	
	Socio-economic status as a child	
	Socio-economic status as an adult	
	Carcinogens	

Figure 1.1. Metadata Fields, Adapted from Daneshvari Berry and Edgar (2021).

The demographics of the decedents of the NMDID are diverse and numerous. With respect to manner of death, 38.6% are accidents, 34.6% natural, 7.4% homicides, 15.4% suicides, and 4% undetermined. Expected deaths (e.g., elderly persons in nursing homes,

and physician-attended deaths) were excluded from the sample. Of the sample, 10,750 are male, 4,475 are female, and 17 are unknown. More than two-thirds of the CT scans have no discernable decomposition (Daneshvari Berry & Edgar 2021). There are also metadata fields for race, tribe, and ethnicity and these fields can be further divided depending on paternal and maternal attribution (ancestry). The metadata field for race includes 17 identification options, the field for ethnicity includes four, and there are 24 tribal affiliation options. These metadata categories are either self-identified or determined by the next-of-kin (NOK) of the decedent.

The NMDID is an invaluable resource for a variety of fields because of the large sample of high-resolution whole body CT scans coupled with thorough documentation. Metadata can be analyzed in isolation for public health purposes, educators can utilize the scans for learning about anatomy and human variation, and more recently, the database is being used to assess "Novel Radiomic and Deep Learning Metrics for COVID-19 Disease." Johns Hopkins School of Medicine researchers are utilizing the database to compare lung tissue from patients with COVID-19 pneumonia with NMDID individuals who have died from relevant causes such as lung cancer (Jones 2020).

Daubert Standard

Expert witness testimony criteria for admissibility have been established and further elucidated byway of three United States Supreme Court decisions (Christensen & Crowder 2009). The 1993 case concerning *Daubert v. Merrell Dow Pharmaceuticals* supplanted the federal court ruling on *Frye* and established that the scientific community had no authority in establishing credibility for scientific evidence, but that judges instead

were the gatekeepers of this decision (Grivas & Komar 2008). In 1999, the decision from *Kumho Tire v. Carmichael*, resulted in the ruling that technical expert testimony, not just scientific, was subject to Daubert's general reliability requirement (Grivas & Komar 2008). The decision of *General Electric Co. v. Joiner* resulted in the ruling that expert testimony cannot be admitted if the data is not sufficiently connected or relevant to the conclusions of the case (Christensen & Crowder 2009). Ultimately, these decisions from *Daubert* trilogy (*Daubert v. Merrell Dow Pharmaceuticals, Inc., General Electric Co. v. Joiner,* and *Kumho Tire Co. v. Carmichael*) were meant to certify that scientific or technical testimony admitted as evidence in federal courts was reliable and relevant (Christensen & Crowder 2009).

The aftermath of the *Daubert* trilogy created ripple effects in most scientific disciplines, forensics included. Recently, there has been an increase in the awareness by forensic anthropologists to critically evaluate some of the techniques and methods commonly used in the field (Christensen & Crowder 2009). Specifically, it is necessary that issues like validation, error rates, and professional standards are incorporated into research, but recent investigation suggests that this is seldom the case (Christensen & Crowder 2009).

This could be due to an intrinsic complexity and difficulty in the field of forensic anthropology to conceptualize methods that implement various traditional scientific methodologies and techniques in conjunction with "less rigorous", subjective observational methods (Christensen & Crowder 2009). Despite this ostensible complexity, Grivas and Komar (2008) indicate that many of the anthropological methods

in practice do meet the admissibility criteria under *Kumho,* making revisions and modifications to current methods redundant (Christensen & Crowder 2009, Grivas & Komar 2008). Although this ruling may recognize experience-based expert testimony from a practicing forensic anthropologist, it could also determine visually-based estimation methods to be too subjective (Christensen & Crowder 2009).

The Scientific Working Group for Forensic Anthropology (SWGANTH) was specifically developed for the purpose of initiating and boosting discourse within the field and creating and publicizing and distributing consensus guidelines for the practice of forensic anthropology (Christensen & Crowder 2009). The subjectivity ad potential error inherent to some estimation methods can be mitigated through quality assurance measures including proper training, accreditation, certification, development of standard operating procedures (Christensen & Crowder 2009), and validation studies of analytical methods to establish method reliability and accuracy (e.g., Curtis 2017, Lewis & Garvin 2016). For the present study, cross-validation rates, accuracies, and intraobserver error rates are presented in chapter 4.

Hypothesis

It is hypothesized that (1) a population-inclusive model for estimating sex will produce classification accuracies that are not statistically different than the population-specific models. Additionally, that (2) a population-inclusive model will perform better than the population-specific models when applied across populations.

Organization of Chapters

The following chapters will demonstrate the utility of a population-inclusive model for estimating sex using volume rendered CT images of the skull. Chapter two presents the relevant previous research that collectively addresses questions of sexual dimorphism and the ways in which current sex estimation methods in forensic anthropology rely on population-specific data in the construction of the biological profile. Additionally, nonmetric and metric sex estimation methods using the crania and post-crania will be examined, with particular attention to the skull. Chapter three outlines the detailed methodology for the construction of the study sample, creation and modification of the CT scans, as well as the statistical methods employed to test this study's hypothesis and evaluate intra-observer error. Chapter four provides descriptive summaries of the analyses run with SPSS as well as the relative performances of the population-inclusive and population-specific models. Chapter five discusses the results for the previous chapter and explores the broader implications of the statistical analyses in relation to the present study as well as standard practice in the field. Chapter six concludes this study by underscoring the initial research rationale as well as the broad implications of an accurate sex estimation that is population-inclusive rather than exclusive.

CHAPTER TWO: PREVIOUS RESEARCH

This chapter discusses the methods and techniques for estimating sex, be that visually, with metric analysis, or Computed Tomography (CT), sexual dimorphism which functions as the primary mechanism that allows for sex estimation, and contextualizes sex estimation in current research and studies which emphasize a need for population-specific (i.e., exclusive) models. Currently, there is an overemphasis on using population-specific models when developing skeletal profiles. Sex estimation models employed by professional anthropologists are population-specific (e.g., Spradley & Jantz 2011; Walker 2008) and can only be applied with the highest accuracy to populations similar to the sample populations. Simply put, the ancestry or population affinity of skeletal remains must be known in order to apply the correct sex estimation methods. Additionally, ancestry estimation methods used presently have not considered proximate mechanisms or foundational theories for biological anthropology and human variation (e.g., evolutionary theory or ecogeographic) in their methodological development, leading some forensic anthropologists to question their predictive value (Bethard & DiGangi 2020). Both of these limitations, and especially cases where ancestry is not known, demonstrate the necessity for population-inclusive (or non-specific) models for estimating sex.

Sexual Dimorphism

When conducting forensic anthropological analyses, the estimation of sex is considered a significant component of the biological profile and is oftentimes one of the first steps undertaken in establishing the biological profile (Spradley & Jantz 2011;

Tallman 2019). Sex estimation is an integral component of this profile because it often serves as the foundation on which other assumptions and identification are made and dictates the methods used to estimate age and stature (Klales 2013; Stewart 1979; Tallman 2019). The estimation of the sex of an unidentified skeleton can also help reduce the search parameters of missing-person databases by approximately 50% (Spradley 2016). Estimating the sex of a skeleton is only possible because of the sexual dimorphism that exists between males and females. These differences are due to a variety of factors that include thermoregulation and biomechanical processes involved in obstetrics, sexual selection, mating preferences, allometric considerations, hormonal fluctuation, and production involved in puberty which ultimately results in an increase in bone growth (Arsuaga & Carretero 1994; Best et al. 2018; Frayer & Wolpoff, 1985; Scheuer & Black 2004:140). Generally considered secondary sex characteristics, the onset of skeletal sexually dimorphic trait expressions, which occur during adolescence, coincides with increased levels of circulating sex steroids such as androgens and estrogens (Stock 2018; Vanderschueren et al. 2004). The steroids that drive sexual maturation play an essential role in skeletal growth and development, specifically because they are active within the cells and tissues of the musculoskeletal system throughout ontogeny and thus contribute to normal growth and development of the skeleton (Bouillon et al. 2004; Carson & Manolagas 2015; Vanderschueren et al. 2004; Vanderschueren et al. 2006; Saggese et al. 2002; Stock 2018). From an evolutionary perspective, proximate explanations for sexual dimorphism suggest that it is due to nutritional stress or overall improvements in the environment during growth and development (Frayer & Wolpoff 1985). Ultimate

causation models indicate sexual dimorphism is a genetic adaptation to different ecological, social, or economic factors, with selection as a primary mechanism (Frayer & Wolpoff 1985).

On average, males tend to have a larger stature, more robust cranial and facial features, and greater muscularity relative to the average female (Frayer & Wolpoff 1985). Most of the sex differences in the human skeleton are not observed in infants, children, or subadults which further indicates that the dimorphism is due to the hormonal events occurring at puberty (Austin & King 2016; Osipov et al. 2013; Frayer & Wolpoff 1985). A preliminary study by Stock (2018) explored the age of full expression of the sexually dimorphic traits of the cranium and mandible. The traits examined included the nuchal crest, mastoid process, glabella, and mental eminence which were measured in a longitudinal sample of lateral cranial radiographs from 10 males and 10 females of European-derived ancestry born from 1931 to 1961 from the Denver growth study. The observed age-at-full-trait-expression was compared with the age upon third molar eruption (i.e., dental maturity) (Stock 2018). Stock (2018) found that the mental eminence reached full expression significantly earlier relative to dental development, albeit with a high degree of variability, and that the age of full expression for the other traits did not differ significantly relative to dental maturity (Stock 2018).

Arsuaga and Carretero (1994) used multivariate analyses to evaluate the sexual dimorphism of the pelvis. Specifically, they examined whether or not the observable metric and nonmetric differences were due to allometric factors (Arsuaga & Carretero 1994). They assessed the pelves of 418 individuals of known sex (227 males; 191

females) who were born in the Beira Litoral region of Portugal and died between 1820 to 1920. The nonmetric traits they studied included the angle of union of the ischiopubic ramus and pubic symphysis, subpubic concavity, preauricular sulcus, greater sciatic notch shape, ventral arc, medial aspect of the ischiopubic ramus, pubic crest robustness, composite arc, shape of obturator foramen, and dorso-symphyseal surface depressions. Arsuaga and Carretero (1994) observed marked sexual dimorphism in the position of the sacroiliac joint as well as the shape of the sciatic notch. Additionally, they examined australopithecine hip bones of unknown sex and noted that the specimens exhibited long pubic bones and iliopectineal lines. Ultimately, Arsuaga and Carretero (1994) conclude that a substantial amount of sexual dimorphism of the hip bone is due to size alone, but that sex-linked morphological and metric differences also exist, though they vary in significance. Additionally, they observed that the modern females in the collection exhibited more variability than the males in nonmetric traits. This suggests that the dimorphism observed is not due exclusively to allometric differences in the sizes. With respect to the australopithecine specimens, Arsuaga and Carretero (1994) conclude that obstetric selective pressures played no role in elongating the pelvic canal of australopithecines. Instead, the evolution and selection for the long pubis and subsequent sexual dimorphism in *Australopithecus* was more likely related to thermoregulation and biomechanics involved in obstetrics (Arsuaga & Carretero, 1994).

In adult skeletons, sexual dimorphism has been observed in the skull, pelvis, and postcranial skeletal elements (Austin & King 2016, Phenice 1969). Because of this, the methods utilized in sex estimation have generally focused on the pelvis, where the

reproductive difference between males and females is distinguishable, and the cranium where the size and morphology are varied and well represented (Işcan 2005).

Gender Identity and Assigned Sex

When employing or discussing methods of sex estimation, research has historically functioned and presently operates under the assumption that sex is binary and fixed. When sex is predicted from the skeleton, the estimation for many people may be correlated with self-identified sex and gender, but not in all cases. For transgender, non-binary, gender non-conforming, gender fluid, or genderqueer individuals who make up an increasing percentage of the forensic record, assigned sex may conflict with their personal gender identity and thus may fail to capture who that individual was in life. Throughout history and presently, there has been a conscious erasure and feigned ignorance of the bodies and biology of individuals who exist outside of a "normative model centered on white cis men" (DuBois & Shattuck-Heidorn 2021:4).

Sex or assigned sex refers to an individual's classification at birth, usually male or female, and is largely based on the visual assessment of external anatomy, but is actually based on a combination of characteristics including chromosomes, hormones, internal and external reproductive organs, and secondary sex characteristics (GLAAD 2021). Gender is distinguished from assigned sex is characterized by shared behaviors, activities, roles, and qualities that society dictates are appropriate for certain genders and may potentially interact with, but remain distinct from, biological sex (Tallman et al. 2021; World Health Organization 2021). Gender roles as well as social constructs vary across societies and evolve over time and when individuals do not fall within

conventional established gender norms they are often exposed to stigmas, discriminatory behavior, and social exclusion (World Health Organization 2021).

Gender identity is defined as an individual's internal sense of their own gender which is not visible to others (GLAAD 2021) and which is contextualized based on socio-cultural and structural experiences and self-expression (DuBois & Shattuck-Heidorn 2021). Cisgender denotes an individual whose gender identity correlates, based on societal constructs, with their assigned sex, or more simply an individual who is *non-transgender* (DuBois & Shattuck-Heidorn 2021; GLAAD 2021). Transgender or trans (shorthand) is an umbrella term that generally refers to an individual whose gender identity does not align with the sex they were assigned at birth (DuBois & Shattuck-Heidorn 2021; GLAAD 2021). Nonbinary, genderqueer, or gender diverse describes an individual whose gender identity lies outside the socially conventional categories of man and woman. Gender non-conforming signifies an individual whose gender expression is not conventionally feminine or masculine (DuBois & Shattuck-Heidorn 2021; GLAAD 2021).

Presently there is a dearth of research regarding individuals who fall outside the conventional male or female sex which highlights a need for a standardized system for assessing and reporting cases with gender diverse individuals (Tallman et al. 2021). Because sex and gender identities intersect variably and are rooted in biocultural, performative processes that move beyond biological processes, it is important for forensic anthropologists to consider the self-identity of individuals in life to avoid restricting identity to biological or anatomical definitions and biological specimens in isolation

(Tallman et al. 2021). Estimating the sex of unidentified remains as indeterminate or ambiguous can signify that (1) the methods produce no resolution, (2) the anatomical evidence is conflicting, (3) there is a lack of skeletal data, (4) or the analyst cannot be certain in the assessment (Tallman et al. 2021). Further, Tallman et al. (2021) conclude that most importantly, forensic anthropologists should not "unjustifiably assume that individuals are cisgender and avoid assumptions that all individuals share the analyst's perspective on sex and gender and associated expressions" (Tallman et al. 2021:15). There is currently no existing standardized method of reporting if a decedent is gender diverse (Tallman et al. 2021). Moreover, current sex estimation methods for analysis and reporting have not evolved to be aligned with modern social perceptions of sex, gender, and identity (Tallman et al. 2021).

Methods of Estimating Sex

The sex of unidentified skeletal remains can be estimated in a variety of ways and with various elements, including visual and metric assessment of cranial and post-cranial remains, as well as with the application of statistical software like Fordisc. The visual and metric methods are usually complementary and tend to result in similar levels of accuracy (Acsád & Nemeskéri 1970; Krogman & Işcan 1986; Rogers 1999). Visual assessment of the *morphology*, or form, of traits and features, which range from gracile to robust, is most easily employed when estimating sex (Frayer & Wolpoff 1985; Spradley 2016; Tallman, 2019). Early methods of sex estimation were based on a gestalt analysis and ordinal scores which were then compiled for a decision table or majority-rule-type approach (Brothwell 1963; Burns 2006; Gray 1966; Hrdlicka 1952; Krogman 1962;

Montagu 1960; Rogers & Saunders 1994; Stewart 1952) without the use of statistical probabilities which current visual methods (Klales et al. 2012; Rennie 2018; Walker 2005, 2008) and Fordisc (Jantz & Ousley 2013) employ. The old methods of gestalt visual assessment can only be highly accurate in the hands of researchers with many years of experience (Phenice 1969). Metric methods, likewise have undergone a transformation since the advent of their use (Letterman 1941; Washburn 1948,1949) to modern studies that employ probabilities and are used in the field currently (Spradley & Jantz 2011). Regardless of the type of analysis being used, nonmetric or metric, many of the current methods used in forensic casework and anthropological research are often based only on populations of U.S. Black U.S. white, and pre-contact Native American skeletons (Tise et al. 2013), many of which come from the Hamann-Todd, Terry, and Bass skeletal collections (Spradley & Jantz 2011), which can perform poorly when applied to morphologically different groups (Roth et al. 2013; Tallman 2019; Tise et al. 2013).

Estimation of Sex from the Pelvis

In forensic and biological anthropology, the pelvis is generally understood as the best indicator of sex (Buikstra and Ubelaker 1994; Klales et al. 2012; Phenice 1969; Walker 2005). The sex differences observable in the pelvis are broadly due to biomechanics involved in obstetrics as well as allometry (Arsuaga & Carretero 1994; Klales 2021). Research on sexual dimorphism in the pelvis has focused on the greater sciatic notch (Letterman 1941; Walker 2005), pubis (Klales et al 2012; Phenice 1969; Washburn 1948, 1949),

Letterman (1941) examined the greater sciatic notch (GSN) in a "comparative study of sexual and racial features" (Letterman 1941:101). In this study, he assessed 426 pelves from U.S. Black and U.S. white skeletons from the eponymous Terry skeletal collection under the direction of Dr. Mildred Trotter. Letterman (1941) discerned metric differences between male and female pelves specifically in the width and height of the GSN as well as distances between the posterior inferior iliac spine (Letterman 1941). While many differences in the pelvis were attributed to sex, Letterman (1941) claimed that race also played a significant role. Walker (2008), discussed more later in the chapter, revisited the GSN to develop probabilities for sex estimation. These analyses and subsequent regression equations were also developed on U.S. Black, U.S. white, and English (British) skeletons.

Washburn (1948) conducted one of the first analyses on the human pubic bone to assess the nature and degree of sexual differentiation in the pelvis. Three hundred adult skeletons from the Hamann Museum of Anatomy and Comparative Anthropology of Western Reserve University, of U.S. Black and U.S. white population affinity, made up the sample (Washburn 1948). The ischium and pubis were measured following methods first outlined by Schultz (1930) in a study involving non-human primates. Washburn's (1948), study sample was made up of the same demographic as Letterman's (1941), who also found metric differences that were influenced by sex and ancestry.

Phenice (1969), later developed an objective method for estimating sex using three traits of the pubis. Specifically, he generated a method that could be used by researchers with limited experience on complete and fragmentary elements. Phenice

(1969) used a sample of 275 adult skeletons from the Terry Skeletal Collection which is comprised exclusively of U.S. Black and U.S. white and examined morphological traits of the pelvis including the ventral arc, subpubic concavity, and medial aspect of the ischiopubic ramus. Phenice (1969) hypothesized that using these three criteria would result in higher accuracy rates when estimating the sex of a skeleton using the pubis. Ultimately, Phenice developed a method with absolute character traits, meaning there were no intermediate variants of trait morphologies. It was his hope that this simple approach would create more objective and accurate results, but its exclusive, population-specific nature prevents this method from being applied to human populations that differ from the original sample groups used in his study.

In *Standards*, Buikstra and Ubelaker (1994) present a modified version of Phenice's (1969) three trait visual method of estimating sex from the pubis as well as a method of estimation using the GSN morphology and preauricular sulcus. Each method of sex estimation using the pelvis incorporates the use of ordinal scores as indicators of "masculinity" (indicating male) and "femininity" (indicating female) (Buikstra & Ubelaker 1994). The pubic estimation method includes the visual assessment of the ventral arc, subpubic concavity, and ischiopubic ramus ridge. Each trait can be scored on a scale from 1 (female) to 3 (male), with an intermediate score of 2 indicating sex as ambiguous. The GSN can be scored on a scale from 1 to 5 with 1 indicating "typical female morphology" and "the higher numbers [showing] masculine conformations" (Buikstra & Ubelaker 1994:18). Lastly, the preauricular sulcus can also be given an ordinal score indicating overall morphology, but in general, the existence of a

preauricular sulcus in any form is thought to indicate female, whereas an absence indicated male (Buikstra and Ubelaker 1994).

Walker (2005), as aforementioned, evaluated the accuracy of current nonmetric methodology used in scoring sex differences, specifically in the greater sciatic notch. Walker (2005) does this using a sample of 296 skeletons of known age and sex. Walker (2005) used museum collections to help establish a range of variation in greater sciatic notch morphology and then applied the technique to a collective sample of U.S. Black, U.S. white, and English skeletons from the Hamann-Todd anatomical collection, the Terry collection at the Smithsonian, and St. Bride's collection at St. Bride's Church in London. This sample had a combined age range of late 18th-century to late 19th-century. Walker (2005) found a strong correlation between age at death and sciatic notch score in both sexes, meaning that individuals who die at a younger age are more likely to have wider, more "feminine" appearing sciatic notches than individuals who die at an older age and that there are sex-related differences. Walker (2005) asserted that the pattern observed in this sample contradicted genetic explanations because of the greater similarity between U.S. white and U.S. Black pelves and the dis-similarity between U.S. white and English pelves. Walker (2005) postulated that vitamin D deficiency (i.e., rickets and osteomalacia) was a non-genetic explanation for the sciatic notch differences between the English skeletons and U.S. skeletal collection.

Klales et al. (2012) critically examined the nonmetric traits first implemented by Phenice (1969). Using a sample of 310 adult left pelves with known ages and sex from the Hamann-Todd Human Osteological Collection and W.M. Bass Donated Skeletal

Collection, Klales et al. (2012) re-examined the ventral arc, subpubic concavity, and the medial aspect of the ischiopubic ramus and modified Phenice's (1969) character states to encompass population variation. The sample population was comprised of mostly U.S. Black and U.S. white males and females, but included one Japanese female, eight "Hispanic" males, one "Mexican" male, and one Asian male (Klales et al. 2012: 107). The analysis included ordinal scoring of traits, analyzing the scores using statistical classification, and testing intra- and inter-observer agreement in order to ensure the highest level of reliability and validity (Klales et al. 2012). Similar to Walker (2005), Klales et al. (2012) blindly scored the sample population using an unbiased five-point scale (without any assumption of masculinity or femininity) and expanded trait descriptions. They looked for significant differences in trait frequencies between the two collections (Freeman-Halton test), estimated the correlative relationship between two observed ordinal variables (polychoric correlations), and employed LR to the scores for each trait. Intra- and inter-observer error was also evaluated using participants with varied experience to test the objectivity of the character trait scoring method and ultimately found that the proposed method is highly reliable for scoring features. The authors also weighed traits differently based on diagnostic and informative condition. The ventral arc was the most reliable for estimating sex (~88% correct classifications) followed by subpubic contour (~86%), and medial aspect of the ischiopubic ramus (~75%). This study ultimately contradicted Phenice's (1969) argument that intermediate forms of character traits do not occur, and importantly established logistic regression equations with probabilities that can be used reliably especially in light of *Daubert*

criteria for scientific standards. Additionally, Klales et al. (2012) found that there were no significant differences between the U.S. Black and U.S. white groups in pelvic morphology and pooled the sample to create a partially population-inclusive method of nonmetric sex estimation.

Karsten (2017) critically examined the incidence and morphology of the preauricular sulcus and how it related to the sex of an individual. The sample was comprised of 500 left pelves from U.S. Black (m=117; f=126) and U.S. white (m=144; f=113) individuals (from the Hamann-Todd Osteological Collection). The preauricular sulci for each individual were scored based on presence and morphology and differential significance between males and females was tested using chi-square methods. Additionally, ancestry was investigated by means of multiple regression analysis. Karsten (2017) then used a stratified random sampling method to select the os coxae to ensure an evenly distributed range of ancestry, age, and sex. Ultimately, Karsten (2017) found that the presence or absence of the preauricular sulcus was a good indicator of sex in 75.8% of individuals; however, there were differences in accuracy between females (89.96%) and males (62.84%). Male pelves were significantly more likely to have short, narrow sulci relative to females. Females, on the other hand, exhibited significantly wider and longer sulci. Karsten (2017) concluded that while the preauricular sulcus should not be considered "positive proof" of female sex, the absence could be used to suggest male sex. Karsten (2017) specified that the preauricular sulcus should be used for estimating sex with extreme caution because the presence of a sulcus cannot conclusively and definitively distinguish male from female. Additionally, Karsten (2017), did not find any

significant relationship between the expression of the preauricular sulcus and ancestry despite "documented variation between populations from around the globe" (Karsten 2017: 607).

Long Bones

On average and in most populations, males will have more robust, rugose muscle attachment sites throughout the body, relative to females (Klales 2020). The appreciable differences between sexes in the pelvis, as well as the positioning and orientation of the pelvis due to obstetrics, result in other sex-based differences throughout the body (Klales 2020). These differences include general size disparities, with males tending to be larger and females smaller, as well as biomechanical distinctions such as the "Q-angle" of the knee and carrying angle of the elbow (Klales 2020).

Rogers (1999) introduced a new method of estimating sex based on four morphological indicators of the posterior, distal humerus. Rogers (1999) hypothesizes that the distal morphology of the humerus is related to the carrying angle of the elbow which is inherently sexually dimorphic. The study sample used to construct the method was comprised of a random sample of males and females from the Grant Skeleton Collection of the University of Toronto's Department of Anthropology, and all but one of these individuals were white. From this initial analysis, the trochlea, olecranon fossa, and medial epicondyle all presented as potentially sexually dimorphic. A blind test was conducted on a hold-out-sample of 20 males and 19 females, utilizing these features and producing an accuracy for estimating sex that ranged from 56 to 86% (Rogers 1999). This technique was then applied to a modern sample from the University of New Mexico

(UNM) of 35 white individuals (28 males and 7 females) who ranged in age from 20 to 90 years. The five features examine during this analysis included the (1) orientation of the medial aspect of the trochlea, (2) trochlear constriction, (3) trochlear symmetry, (4) olecranon fossa shape and depth, and (5) angle of the medial epicondyle. The modified technique from the UNM sample was then applied to another modern sample of 93 white skeletons (74 male and 19 female) with an overall accuracy of 92%. Rogers (1999) concludes that because the method relies on shape instead of size, the likelihood that the technique could be applied across populations is greater than if it focused on size. Vance et al. (2011) later modified this method to include a 5-graded ordinal scoring system that categorized individuals as either "male" (scores of 1 or 2), "female" (scores of 4 and 5), and ambiguous (scores of 3).

Rogers et al. (2000) conducted a study evaluating the reliability of the rhomboid fossa of the clavicle as an indicator of sex and age. The study sample was comprised of 344 individuals (113 females and 231 males) from the William F. McCormick collection at the University of Tennessee in Knoxville. Due to the "regional demographics of the collection", the sample was predominately white individuals with only 38 Black skeletons and 5 "'other' non-whites" included (Rogers et al. 2000: 62). The results of this study indicate that the presence of a rhomboid fossa likely indicates the sex is male, but if the costoclavicular attachment "exhibits an impression, a tubercle, or leaves no trace" (Rogers et al. 2000: 66) then the sex cannot be estimated.

As abovementioned, Vance et al. (2011) adapted previous studies (e.g., Rogers 1999) that documented sex-specific morphological features of the distal humerus, to include a

5-graded ordinal scoring system. The study sample was comprised of 608 humeri (420 belonging to males and 188 belonging to females) from the Pretoria Bone Collection at the Department of Anatomy, University of Pretoria and Raymond A. Dart Collection at the University of Witwatersrand. The sample included a random distribution of white and Black South Africans of all socioeconomic backgrounds (Vance et al. 2011). The methodological system included designations of "clearly male" (scored as 1), "cautiously male" (scored as 2), "ambiguous" (scored as 3), "cautiously female" (scored as 4), and "clearly female" (scored as a 5) (Vance et al. 2011:711). The three traits – olecranon fossa shape, angle of the medial epicondyle, and trochlear extension – would each be given a score from 1 to 5, and the total combined numerical value was used to make an overall estimation of the sex. Total scores between 3 and 9 were estimated to be male, a total score of 9 estimated to be ambiguous, and scores between 10 and 15 estimated to be female (Vance et al. 2011). Despite producing lower accuracies than previous studies, Vance et al. (2011) came to similar conclusions: that the distal humerus is sexually dimorphic and can be used as a reasonably reliable indicator of sex if no other remains are available.

Forensic anthropologists employ Fordisc when estimating the ancestry (population affinity), stature, and sex of unidentified skeletal remains (Jantz & Ousley 2013). Fordisc is a statistical software that performs linear discriminant function analysis (LDFA) and linear regression analyses with prediction intervals from bone measurements (Jantz & Ousley 2013). The classification functions of Fordisc are based on a database of skeletal measurements from the Forensic Data Bank (FDB), which is comprised of skeletal

information from forensic cases as well as museum and university collections (Jantz & Ousley 2013). The FDB also includes added information for the individuals that make up the skeletal collection including place of birth, medical history, occupation, stature, and weight, in addition to skeletal data like cranial and postcranial metrics, suture closure information, aging criteria scores, nonmetric trait observations and congenital traits (Jantz & Ousley 2013). Fordisc is a useful tool for the professional anthropologist for several reasons. Firstly, the statistical software calculates classification functions based on any number of measurement combinations available as well as any combination of postcranial measurements which enables estimates of sex and ancestry to be made from complete and incomplete remains (Jantz & Ousley 2013). Fordisc also calculates numerical results including "cross-validated classification accuracies, tests of multivariate assumptions and posterior and typicality probabilities, which are essential in evaluating classification performance and individual classifications" (Jantz & Ousley 2013:97). Researchers are also able to view the statistical results as graphs which can help to demonstrate classifications and types of multivariate variation (Jantz & Ousley 2013). Fordisc additionally has the capacity to classify unidentified individuals using the included worldwide craniometrics database from both Howells and 19th-century American samples (Jantz & Ousley 2013). During statistical analyses, the software includes a stepwise selection of measurements which shows the most valuable measurements in regard to classification (Jantz & Ousley 2013). Most importantly, Fordisc can also be updated through the internet which enables researchers to have access to the most up-to-date information, including modern samples and comprehensive

population groups (Jantz & Ousley 2013). Ultimately, Fordisc is a useful tool for forensic anthropologist because it employs validated and familiar statistical procedures in a user-friendly and adaptable form for efficient analysis of a single case (Jantz & Ousley 2013).

Sex Estimation Using the Skull

When the pelvis or long bones are unavailable for sex estimation, the skull can be used with statistical accuracy to either through visual inspection of morphological features or through metric measurements (Buikstra & Ubelaker 1994; Langley et al. 2016). In cases where remains are incomplete, such as seen in outdoor contexts, taphonomic processes inhibit a complete recovery of all skeletal elements (Spradley & Jantz 2011). Historically, the cranium was considered the second-best indicator of sex following the pelvis (Bass 2005; Pickering & Bachman 1997), but more recent studies suggest that some metric measurements of the long bones, such as the femur, are a better indicator (Albanese et al. 2008; France 1998; Spradley & Jantz 2011; Patterson & Tallman 2019). Nonetheless, the skull has provided highly accurate results and is therefore regularly utilized in conjunction with pelvic assessments or when pelvic traits are unavailable (Bass 2005; Krogman & Işcan 1986; Stewart 1979).

Methodological advancements as well as an increase in pressure for anthropologically-based fields to become more “scientific” and “objective” have resulted in statistically-based osteometric techniques supplementing more subjective and visually-based methods. Regardless, visually-based estimation techniques continue to persist alongside more statistically based methods. This is attributed to preconceived notions that osteometric data collection is more time consuming and usually requires expensive,

specialized anthropometric equipment (Walker 2008). Additionally, there is an inherent difficulty in establishing measurements that effectively capture the subtle visual variations in sexually dimorphic shapes and features (Graw et al. 1999). Poor preservation and subsequent small sample sizes also impede otherwise strong statistical techniques that would require normally distributed data (Walker 2008).

Metric Traits of the Skull

Metric methods and techniques typically involve the analysis of skeletal measurements as well as multivariate shape analyses (Lewis & Garvin 2016). While it is generally argued that the visual, morphological examination of the skull, which involves assessing robusticity or gracility of specific features like the supraorbital ridge or mastoid without traditional measurements, can still produce correct classification better than metric analysis, there is a lack of research to support this (Spradley & Jantz 2011). Currently, forensic anthropologists generally employ craniometrics methods of sex estimation accompanied by statistical analyses with produced probabilities (Dayal et al. 2008; Jantz and Ousley 2005; Spradley and Jantz 2011).

Dayal et al. (2008) employed discriminant function analysis to assess the accuracy of estimating sex on a population of South Africans. The study sample was comprised of 120 crania of Black South Africans (60 males and 60 females) from the Raymond Dart Collection housed in the School of Anatomical Sciences, University of Witwatersrand, Johannes burg, South Africa, who's age at death ranged from 25-70 years. For this study, the skeletal remains of the individuals were classified as Zulu, Sotho, Tswana, and Xhosa

(Dayal et al. (2008). The measurements and descriptive definitions used in the study were derived from Martin and Knussman (1988) (Figure 2.1).

No.	Measurement	Abbreviation	Definition
1	Maximum cranial length[b]	g-op	Glabella to opistocranion
2	Basi-bregmatic height[b]	ba-b	Basion to bregma
3	Minimum frontal breadth[b]	ft-ft	Frontotemporale to frontotemporale
4	Bizygomatic breadth[b]	zy-zy	Zygion to zygion
5	Basion–nasion length[b]	ba-n	Basion to nasion
6	Basion–nasospinale length[b]	ba-ns	Basion to nasospinale
7	Basion-prosthion length[b]	ba-pr	Basion to prosthion
8	Nasal height[a]	n-ns	Nasion to nasospinale
9	Nasal breadth[a]	NLB	Maximum width of apertura piriformis
10	Orbital breadth[a]	mf-ek	Maxillofrontale to ectoconchion
11	Orbital height[a]	OBH	Maximum height from the upper to lower orbital borders perpendicular to the mf-ek line
12	Palate breadth[b]	ekm-ekm	Ectomolare to ectomolare
13	Palate length[b]	ol-sta	Orale to staphylion
14	Upper facial height[a]	n-av	Nasion to alveolare
15	Total facial height[a]	n-gn	Nasion to gnathion
16	Bicondylar breadth[a]	kdl-kdl	Condylion laterale to condylion laterale
17	Bigonial breadth[a]	go-go	Gonion to gonion
18	Ramus height[c]	rh	Gonion to uppermost part of ramus
19	Ramus breadth[a]	rb	Minimum breadth between the anterior and posterior borders of the ascending ramus
20	Total mandibular length[c]	ml	Gnathion to perpendicular line between the condylar processes
21	Gonion-gnathion length[c]	go-gn	Gonion to gnathion

Variable labels generally follow Martin's system, except where the system gives no label, three-letter labels introduced by Howells were used (see Martin and Knussmann, 1988 for details).
[a]Sliding caliper.
[b]Spreading caliper.
[c]Mandibulometer.

Figure 2.1. Measurements and Definitions from Dayal et al. (2008) (Adapted from Martin and Knussman (1988)).

The results of the study showed that all the male measurements, except orbital height, were significantly greater ($p < 0.05$) than the corresponding female measurements. The facial height, bizygomatic breadth, and mandibular ramus height individually produced the highest rates of correct classification. Overall an average accuracy of 80.8% was obtained using the Stepwise analysis for cranial measurements and 85% for the mandibular measurements, including the bigonial breadth, mandibular ramus height, and total mandibular length. The results indicate that all the cranial and mandibular measurements used in the study demonstrated significant sexual dimorphism except for the orbital height; however, the measurements are not individually useful in estimating sex.

In a study by Spradley and Jantz (2011), the hierarchy of effectiveness for estimating sex from cranial and postcranial elements was tested. The objective of the study was to test the theory that the skull is second to the pelvis in its effectiveness at estimating sex and to provide updated classification rates for estimating sex from the postcranial skeleton for forensic anthropologists to implement. Spradley and Jantz (2011:289) assessed a recent forensic sample from the U.S. in order to determine, by element, a "hierarchy of sexing reliability". This sample came from the Forensic Anthropology Data Bank (FDB), which is considered a population-specific dataset for the U.S. because its data is derived from individuals that reflect the population it is currently being applied to (Spradley & Jantz 2011). Spradley and Jantz (2011) were limited in their sample to Black and white American individuals because the sample size for other population groups (e.g., Hispanic, East Asian, and Native American) were too small to obtain meaningful

results (Spradley & Jantz 2011). Additionally, the sample only included adults, 18 years or older, that were born in 1930 or later. This exclusionary parameter was implemented to account for secular change in U.S. populations and in order to include an age range of individuals that was reflective of current forensic anthropological cases (Spradley & Jantz 2011). In order to assess the hierarchy of element reliability, Spradley and Jantz (2011) used measurements usually taken from standard landmarks and analyzed with discriminant function analysis, including 24 cranial, 10 mandibular, and 44 postcranial measurements, and utilized univariate and multivariate methods to statistically assess sex estimation for the mandible, cranial, and postcranial elements. Spradley and Jantz (2011) found that there is a significant difference in sex and ancestry between Black and white American crania, mandible, and postcranial skeleton measurements, except the calcaneus. For Black and white Americans, the cranium provided an overall cross-validated classification rate of 90-91%, whereas numerous postcranial elements provided a cross-validated classification rate between 92% and 94% (Spradley & Jantz 2011).

Morphological Traits of the Skull

Despite an ever-expanding myriad of techniques used to assess sexual dimorphism, biological and forensic anthropologists continue to depend on more traditional methods of visual and sometimes size-related assessments when examining cranial traits for estimating sex (Acsádi & Nemeskéri 1970; Buikstra & Ubelaker 1994; Garvin et al. 2014; Vollner et al. 2009; Walker 2008). Morphological or nonmetric methods of trait assessment often complement the metric analyses and are not based on size, but instead the degree of expression of sexually dimorphic traits (Lewis & Garvin 2016).

Additionally, nonmetric methods continue to be employed because they can readily be applied to fragmentary remains, require no specialized equipment, are relatively rapid, and require minimal experience when compared with metric analyses (Lewis & Garvin 2016).

Morphological sexually dimorphic cranial traits were first identified in 1875 by Paul Broca (Acsádi & Newmeskéri 1970; Buikstra & Ubelaker 1994; Garvin et al. 2014) and have since been examined in numerous studies (Acsádi & Newmeskéri 1970; Buikstra & Ubelaker 1994; Garvin et al. 2014; Graw et al. 1999; Krogman & Işcan 1986; Tallman 2019; Tallman & Go 2018; Walker 2008; Walrath et al. 2004). The cranial trait scoring method presented by Buikstra and Ubelaker (1994) and Walker (2008) is considered the most common cranial nonmetric sex estimation method and is currently utilized by biological and forensic anthropologists (Lewis & Garvin 2016).

The cranial traits most often utilized in nonmetric sex estimation include the supraorbital ridge/glabella, supraorbital margin, mastoid process, mental eminence, and nuchal crest (Buikstra & Ubelaker 1994; Garvin et al. 2014; Walker 2008), and formally documented as early as Broca (1875) and Acsádi and Nemeskéri (1970) who were among the first researchers to construct a sex estimation that incorporated scoring the traits (Klales 2020). Although studies show that all these traits are generally universally sexually dimorphic (Acsádi & Nemeskéri 1970; Buikstra & Ubelaker 1994; Garvin & Ruff 2012; Lewis & Garvin 2016; Walker 2008) overlap between trait expression can occur due to ambiguous expression, age variables, population variation, reduced sexual dimorphism, and variable biomechanics (Klales 2020). For example, variation in mastoid

process size is often attributed to body size, muscle mass, and overall activity because larger muscles necessitate a larger attachment site (Klales 2020). Furthermore, because there will always be males and females that fall variably on the gradient of human variation and sexual dimorphism, the goal of the ordinal scale is to provide a simple and objective method of scoring that relies on assessment of robusticity/gracility or size/shape of a specific trait without any presumption of sex (Klales 2020).

The nuchal crest, which is comprised of the external occipital protuberance, is delimitated as the median line or crest between the right and left nuchal musculature with a highly variable range of expression (White et al. 2012:72). The crest projects from the external occipital protuberance to the posterior border of the foramen magnum (at opisthion) and anchors the nuchal ligament (White et al. 2012:72). In an example of a female nuchal crest, "the external surface of the occipital is smooth with no bony projections visible when the lateral profile is viewed" (Buikstra & Ubelaker 1994:19). In an example of a male nuchal crest or maximal expression, there would be a "massive nuchal crest that projects a considerable distance from the bone and forms a well-defined bony ledge or hook" (Buikstra & Ubelaker 1994:19).

The mastoid process is described by Petaros et al. (2015) as a prominent, tapering bony projection located on the temporal bone, posterior and inferior to the external auditory meatus. The mastoid process forms after birth and grows directly from the flat mastoid portion of the temporal bone (Scheuer & Black 2004:92). The many muscles that attach to this bony prominence, such as the sternocleidomastoid, splenius capitis, and longissimus capitis, exert biomechanical forces that influence the growth and

development of its size and shape (Lieberman 2011; Standring 2008; White et al. 2012:67). These muscles aid in the function (extension, flexion, and rotation) of the head and in some humans and fossil hominids, the presence of a supramastoid crest, the mastoid process can be indicative of the attachment of temporalis muscle as well (White et al. 2012:67). Additionally, the observable morphological differences between sexes and individuals are attributed to variance in the developmental patterns of mastoid cells, which are responsible for the hollow spaces in the interior of the mastoid (Cinamon 2009; White et al. 2012:67). The mastoid is often considered one of the most robust and sexually dimorphic features of the skull (Wahl & Graw 2001; Walrath et al. 2004; Noren et al. 2005; Williams & Rogers 2006; Walker 2008; Saini et al. 2012; Garvin et al. 2014). The visual assessment of the mastoid for estimating sex involves descriptors that document the character or appearance of the process, such as its relative length, volume, and overall robustness or gracility. In females, it might present as a "very small mastoid process that projects only a small distance below the inferior margin of the external auditory meatus and the digastric groove" (Buikstra & Ubelaker 1994:19). In males, the mastoid might present as "massive" with "lengths and widths several times that of the external auditory meatus" (Buikstra & Ubelaker 1994:19).

The supraorbital ridge/glabella or brow ridge of the skull is structurally significant within physical anthropological research due to its anatomical location, morphological variability, and tendency toward gracilization throughout human evolution (Nowaczevska et al. 2014). The supraorbital ridge is located on the frontal bone which at birth is composed of two symmetrical halves, separated initially by the metopic suture (Scheuer

& Black 2004:108). The glabellar region often joins the superciliary arches over the orbits if particularly pronounced (White et al. 2012:61). The frontal sinus is a chamber located in the medial part of the supraorbital ridge, within the frontal bone, sometimes extending throughout the frontal squama, and is continuous with the paranasal sinus (Camargo et al. 2007; Moss & Young 1960; Sardi et al. 2018). The frontal sinus starts its development *in utero*, potentially as a mucosal evagination at the anterior end of the middle nasal meatus or as a derivative of the anterior ethmoidal cells (Scheuer & Black 2004:111). Analysis of the frontal sinus is significant in forensic identification due to its irregular and individual shape which can be used for positive identification (Camargo et al. 2007; da Silva et al. 2009). Additionally, the frontal sinus morphology is variable and constrained by the surrounding structures (Sardi et al. 2018).

A study by Sardi et al. (2018) assessed the covariation with the frontal sinus and surrounding bony structures, specifically looking at the sinus volume and glabellar region, upper nasal region, bone thickness, and endocranial size in a human sample of 149 individuals from Argentina. Sardi et al. (2018) found that frontal sinus volume was correlated with bone thickness measurements from the glabellar region, meaning the sinus tended to be significantly larger and more expanded in individuals with thicker frontal bones and subsequently thicker glabella typically seen in males (Sardi et al. 2018). Nowaczevska et al. (2014) examined the morphological relationship between the glabella and superciliary arch. The results of this research indicated that the glabellar region and superciliary arch experience separate influences during formation, but sex was the strongest overall influence during development (Nowaczevska et al. 2014). Males tended

to have a more curved and projecting nasoglabellar profile, "forming a rounded, loaf-shaped projection that is frequently associated with well-developed supraorbital ridges" (Buikstra & Ubelkaer 1994:20) compared with female crania (Nowaczevska et al. 2014) that have a smooth frontal bone, "with little or no projection at the midline" (Buikstra & Ubelaker 1994:20). The authors attributed this variation to sex differences, as a result of sexual dimorphism, with regard to maturation and spatial reorganization of the cranial structures during ontogeny (Nowaczevska et al. 2014).

Variation in the supraorbital ridge is thought to be due to mechanical stresses related to eating. There are two hypotheses that address the mechanism of morphological variation found in the supraorbital region including the *masticatory-stress hypothesis* and the *structural hypothesis*. The masticatory-stress hypothesis, which posits that variation in the supraorbital region is due to resistance and stress endured during mastication and incision (Endo 1970; Russell 1985; Hiloowala & Trent 1988), has received less support from the anthropological community compared to the structural hypotheses. The structural hypothesis, which includes the craniofacial size model (Lahr & Wright 1996) and spatial model (Moss & Young 1960; Weidenreich 1941). The craniofacial size model posits that the morphological variation in the supraorbital region is directly related to overall craniofacial size (Vinyard & Smith 1997, 2001). Lahr and Wright (1996) found that there was an association between the degree of morphological expression and overall craniofacial size. Vinyard and Smith (2001) likewise observed a correlation between metric traits of the brow ridge and craniofacial size. The spatial model hypothesizes that the structural separation that occurs during ontogeny between the orbital and cerebral

regions of the frontal bone in order to maintain the overall integrity of the skull (Moss & Young 1960; Weidenreich 1941). Despite these hypotheses, the relationship of cranial size and shape and its relationship to the morphology of the brow ridge in modern humans has not been sufficiently explored.

The supraorbital margin forms the upper orbital edges and can occasionally be perforated by a supraorbital notch or foramen (White et al. 2012:61). In males, the supraorbital margin may be a more "thick, rounded margin with a curvature approximating a pencil" (Buikstra & Ubelaker 1994:20). In females, the supraorbital margin may have borders that "feel extremely sharp, like the edge of a slightly dulled knife" (Buikstra & Ubelaker 1994:20). The "sharp" morphology of the supraorbital margin often associated with females was first documented by Broca in 1875. Graw et al. (1999) stipulated that the important indicator of sex using the supraorbital margin is whether the margin presents as an evenly rounded trough shape in a cross-section or whether the inner aspect is angular along the whole of its length, the former indicating male and the latter indicating female. Graw et al. (1999) explored the efficacy of the supraorbital margin as an avenue of sex estimation from the skull in modern humans. The study sample was comprised of 108 skulls (41 females and 67 males) with a date of death between 1964 and 1994 and an even distribution of age at death from the southwest of Germany (Graw et al. 1999). The results of this study indicated that the method was highly reliable with an accuracy rate of approximately 70% (Graw et al. 1999). Garvin and Ruff (2012) evaluated a method to quantify the morphological variation in the skeletal brow ridge (and chin) by utilizing 3D surface laser scans. Their sample consisted

of 119 crania and mandibles of Black and white U.S. males and females from the Terry collection (Garvin & Ruff 2012). Garvin and Ruff (2012) found that sex and ancestry played a substantial role in morphology consistent with the qualitative descriptions traditionally employed. A test of intraobserver error also indicated that the method was highly repeatable (Garvin & Ruff 2012). Furthermore, this study provides a method of morphological assessment of discrete cranial traits that is objective and statistically driven.

The mental eminence, also referred to as the bony chin or mental protuberance, is most commonly defined as the bone protruding anteriorly beyond the mandibular alveolus (Garvin & Ruff 2012). Generally, definitions for this region, as well as the physical borders, have been vague, but usually include the symphyseal eminence (mental osseum), lateral tubercles, and associated mental (incisive) fossae (Garvin & Ruff 2012; Weidenreich 1936). The mental eminence is altered significantly during childhood (Scheuer & Black 2004:140) and during growth and development, the eminence becomes demarcated within the mandibular body by the alveolar margins of the incisors and the "mental sulcus" or incurvation in the bone (White et al. 2012:93). After the eruption of the incisors, the depth also rapidly increases followed by subsequent growth and development to make room for the rest of the dentition (Scheuer & Black 2004:140). Research suggests that the variation in the degree of trait expression does not occur on a simple gradient, but instead is influenced by a combination of traits (Lewis & Garvin 2016). Garvin and Ruff (2012) and Garvin et al. (2014) suggest that the range in chin/mental eminence morphology is due to variation in incorporation of different

characteristics like the prominence of the lateral tubercles, surface area occupied by the mental eminence, projection of the central/midline eminence, and angle of corpus (mandibular body) (Lewis & Garvin 2016). Lewis and Garvin (2016) suggest that the sexually dimorphic differences of the mental eminence may be best understood by assessing the multiple components of the mandible and potentially scoring them independent of one another in order to mitigate observer biases. Some morphometric analyses indicate that the mental eminence is not as sexually dimorphic as originally presumed and, with potential diagnostic accuracy around 75%, should be employed with caution or altogether avoided (Lewis & Garvin 2016). The difference in efficacy regarding sex estimation between the mandibular and the cranial traits could be in part due to the mandible being the only skeletal element, except for the middle ear ossicles, that experiences independent movement (Scheuer & Black 2004:135).

The observable cranial sexual dimorphism is often attributed to a longer and more intense interval growth rate experienced by males, which ultimately results in the relative and absolute larger male skulls and morphological features (Tallman 2019; Viðarsdóttir et al. 2002). With size-related examination, rather than metric assessment, extreme scores enable the possibility of more accurate sex estimation (Ubelaker & DeGaglia 2017). In *Standards*, Buikstra and Ubelaker (1994) presented the five sexually dimorphic cranial traits as part of a method originally devised by Philip Walker to be a guide for practitioners to use in their interpretation of trait expression with the goal of creating comparable results (Lewis & Garvin 2016). The method involves assigning an ordinal score between 1 and 5 through a visual and physical comparison of the traits of the skull

to the standardized descriptions and illustrations (Buikstra & Ubelaker 1994). Traits with more gracile or typically "feminine" appearances receive a score of 1 or 2, ambiguous features receive a score of 3, and more robust or "masculine" features receive a score of 4 or 5 (Buikstra & Ubelaker 1994). A score of "1" is ultimately estimated as female, 2 as probable female, 3 as ambiguous sex, 4 as probable male, and 5 as male (Buikstra & Ubelaker 1994).

Walrath and coauthors (2004) tested the reliability of the visual assessment of cranial traits for estimating sex in an archaeological population of unknown sex. Ten cranial traits were visually assessed (Walrath et al. 2004). These traits included the glabella, mastoid, nuchal plane, zygomatic process of the temporal, superciliary arches, frontal and parietal eminences, external occipital protuberances, zygomatics, frontal profile, and orbital form (Walrath et al. 2004). The sample was comprised of 42 well-preserved, complete, or mostly complete Inuit skulls dating from the Birnirk period (AD500-900) from the University of Pennsylvania Museum (Walrath et al. 2004). This sample was chosen because it represents a homogenous group of individuals, consisting of robust, cold-adapted individuals, which ensured that the reliability of the visual methods could be tested without the confounding variable of inherent population heterogeneity (Walrath et al. 2004). This study employed a multifaceted sex estimation which was a combination of two different standardized methods that were developed during symposia on technique consistency in skeletal biological analyses, including (1) a method from Ferembach et al. (1980) developed during the 1970s to define uniform methods of estimating sex; and (2) a method from Buikstra et al. (1994), developed in a 1991 seminar at the Field Museum

of Natural History in Chicago in response to the Native American Graves Protection and Repatriation Act (NAGPRA) (Walrath et al. 2004 The study from Ferembach et al. (1980) affords researchers precise guidelines and numerical data in order to estimate sex. It includes brief character descriptions with numerical values that range from -2 to +2, which correspond to *hyperfeminine* and *hypermasculine*, respectively (Ferembach et al. 1980; Walrath et al. 2004). The final estimation is a product of weighted averages (Ferembach et al. 1980). Buikstra et al.'s (2014) method, used in tandem with the Ferembach et al.'s (1980) method, contributes more detailed descriptions for each character trait as well as comparative illustrations to evaluate alongside the specimen. Interobserver variation was also evaluated and the results indicate that there was significant interobserver reliability in all the traits except for the frontal and parietal eminences which were non-significant, and although most traits were considered significantly reliable they all varied in the degree of reliability (Walrath et al. 2004). The clarity of trait definitions was essential for accurate and effective estimation of sex via the visual assessment method (Walrath et al. 2004). The authors suggest that mitigating variation between observers could be accomplished through the use of fewer character traits, but with higher levels of detailed descriptions, would improve interobserver reliability (Walrath et al. 2004). In this study, the traits that possessed the clearest definitions and therefore the highest level of agreement, were the glabella, mastoid process, superciliary arches, external occipital protuberance, and zygomatics (Walrath et al. 2004). Ultimately, the results of this study suggest that in visual assessment methods,

precise and clear definitions have a larger influence on interobserver agreement when it comes to sex estimation than number of character traits assessed (Walrath et al. 2004).

Rogers (2005) similarly conducted a study to assess the credibility of morphological sex estimation with particular regard to the Daubert and Mohan criteria for admissibility in a court of law. The Daubert criteria dictates that for a theory or technique to be admissible in court it must (1) be subjected to the scientific method of testing, (2) be published and subjected to peer review, (3) have known or potential error rates, and (4) be generally accepted in the relevant scientific community (which comes from The *Frye* standard) (Edmond & Roach 2011:351). The *Mohan* criteria focuses on evidence presented by an expert witness and determines evidence to be admissible as long as it is "(1) relevant, (2) necessary to assist the trier of fact, (3) caught by an exclusionary rule, and (4) proffered by a properly qualified expert" (Edmond & Roach 2011:381).

Rogers (2005) examined the accuracy and precision of 17 of the most common morphological features of the skull used to estimate sex of unknown skeletal remains. The sample was comprised of 245 skulls from the 19th-century St. Thomas Anglican Church Cemetery in Belleville, Canada. A subset of 46 skulls which were identified based on coffin plates and therefore had known biological profiles, were used to form the basis of the analysis (Rogers 2005). The traits examined in this analysis were based on a suite of 13 traits from Krogman (1955) as well as 4 additional characteristics recommended by various authors (Bass 1995; Buikstra & Ubelaker 1994; Burns 1999; Byers 2002; Roberts 1996; White 2000), and generally focused on the size and architecture of the trait. The 17 traits scored included (1) the general size/architecture of

the skull, (2) forehead, (3) frontal eminence, (4) supraorbital ridge, (5) orbits, (6) nasal aperture, (7) nasals, (8) malars, (9) zygomatic, (10) parietal eminences, (11) mastoid, (12) occipital, (13) occipital condyles, (14) palate, (15) mandible, (16) chin, and (17) tooth size (Rogers 2005).

Rogers (2005) produced an intra-observer error of 12.2% using the combined 17 traits in estimating sex which was only 2% higher than the accepted level of 10%, suggesting that consistency in scoring was difficult. The nasal aperture, zygomatic extension, malar size/rugosity, and supraorbital ridge provided the most accurate estimations, followed by the "chin form" or mental eminence, and nuchal crest (Rogers 2005). Tertiary to the chin and nuchal crest was the mastoid size, and subsequently, the mandibular symphysis/ramus size, forehead shape, and lastly the palate size (Rogers 2005).

Each trait had variably reliable in estimating sex when used in isolation and the accuracies were as follows: general size/architecture of the skull (38%), forehead (44.5%), frontal eminence (31.9%), supraorbital ridge (60.9%), orbits (43.6%), nasal aperture (76.6%), nasals (52.8%), malars (68.4%), zygomatic (70.3%), parietal eminences (28.9%), mastoid (44.7%), occipital/nuchal crest (53.3%), occipital condyles (14%), palate (36.6%), mandible (51.1%), chin form/mental eminence (56.3%), and tooth size (10.3%) (Rogers 2005).

The overall accuracy, when comparing estimated sex with the known sex of the individual was 89.1%. This accuracy was achieved when all 17 morphological features were used in conjunction with one another and when each trait was weighted equally. The trait that had the highest individual accuracy was the nasal aperture (76.6%), no

combination of any two traits exceeded the accuracy of the complete trait list, and only one combination of three traits, including the Zygomatic extension, malar size/rugosity, and nasal aperture, produced an accuracy of 91%. The varying accuracies suggest that no single trait is as accurate in estimating sex as multiple traits used in combination with one another, but that both too few and too many traits can increase observer error by introducing uncertainty into the assessment (Rogers 2005).

Walker (2008) updated the technique put forth by Buikstra and Ubelaker (1994) and tested the accuracy of the sex-specific skull traits to provide a less subjective method of sex estimation that was supported by statistical analyses (logistic regressions). Walker (2008) reevaluated the five sexually dimorphic traits of the skull, including the nuchal crest, mastoid process supra-orbital margin, supra-orbital ridge/glabella, and the mental eminence (Figure 2.2). Using a sample of 304 skulls of known age and sex, mixed ancestry, and historic and modern individuals from the Hamann-Todd Collection at Cleveland Museum of Natural History, the Terry collection at the Smithsonian Institution, and the Stain Bride's Church in London, England, Walker (2008) formed a study group which represented a "modern sample" and an archaeological Native American sample. Univariate statistical analyses were used to calculate the trait-by trait probabilities of estimating sex. In the modern sample, the mastoid process and glabella ranked the highest accuracy, and the nuchal area and orbital margin ranked the worst accuracies. In the Native American sample, the nuchal area was the least sexually dimorphic and the glabella, and mastoid process were the most sexually dimorphic. Multivariate statistical models (binary logistic regression equations) were used to

establish differences among individuals of known sex in order to apply these differences as a model for assigning sexes to individuals with unknown sex. In modern samples that used all five traits to classify sex, the average percent of individuals accurately assigned a sex was 89%. The fewer traits used resulted in a decreased accuracy. Ultimately, Walker (2008) found that after using the standards for sex estimation outlined by Buikstra and Ubelaker (1994), 88% of the modern skull collection were classified correctly and 78% of the Native American collection corresponded with the pelvic estimation of sex. This suggested that Buikstra and Ubelaker's (1994) cranial trait scoring system could be used to accurately assess sex by researchers with varying levels of experience as long as population-specific standards are used. Despite this conclusion, Walker (2008) did conclude that nonmetric sex estimation standards developed for one population should be cautiously applied to other populations to compensate for population variation.

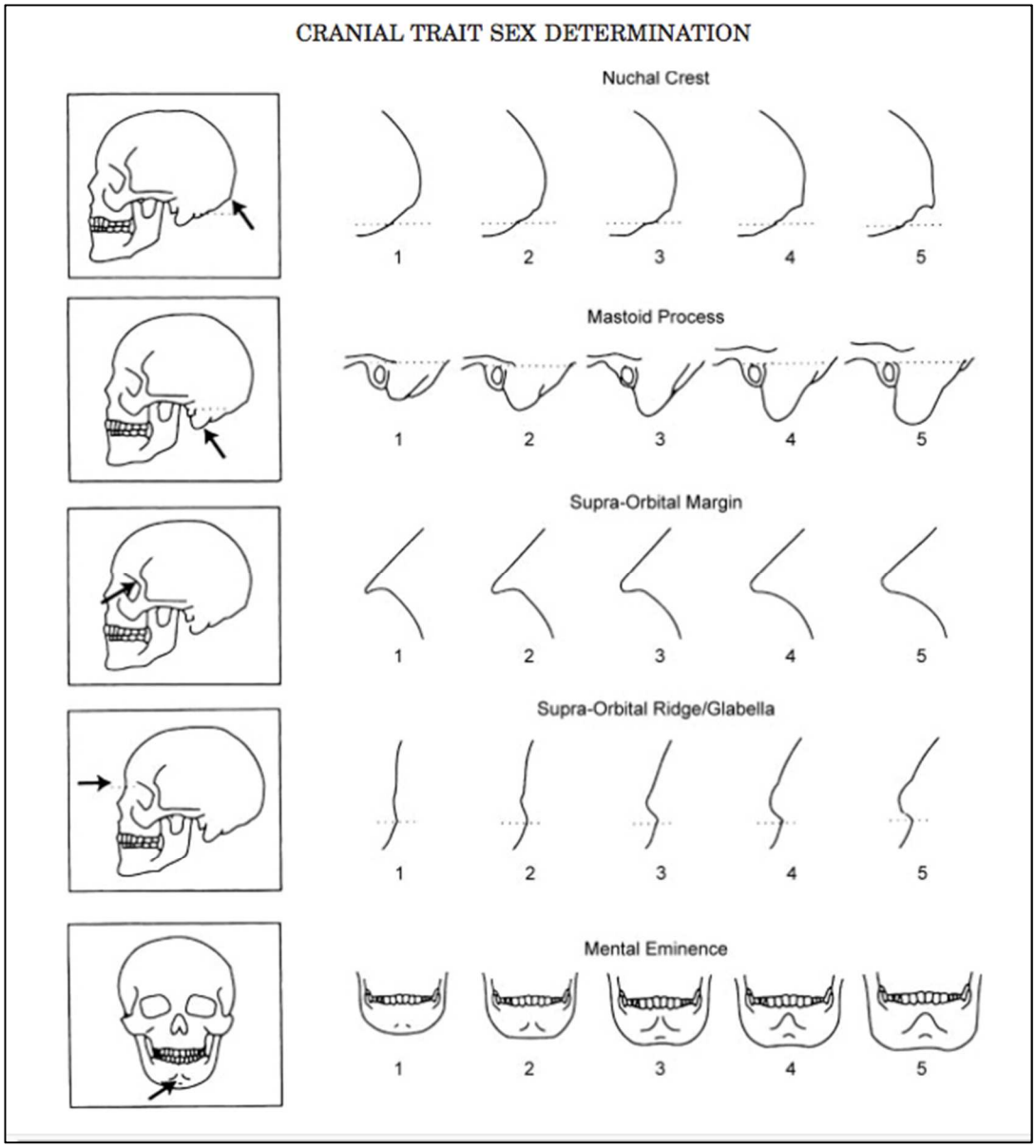

Figure 2.2. Nonmetric Traits (Buikstra & Ubelaker 1994:20 and Walker 2008:41).

Though nonmetric sex estimation is the cornerstone in the development of a biological profile, it faces criticism from within and outside the field due to a perceived lack of testing and validation of qualitative methods of cranial sex determination (Rogers

2005). Despite this criticism, some validation studies exist. Lewis and Garvin (2016) evaluated Walker's (2008) cranial nonmetric method for sex estimation, specifically to assess the reliability and accuracy of the method. The sample was comprised of an approximately even distribution of 135 adult U.S. Black and white females and males from the Hamann-Todd Skeletal Collection (Lewis & Garvin 2016). A modern skeletal sample was specifically not used in this study in order to combat secular change bias from influencing the results (Lewis & Garvin 2016). All five traits (the nuchal crest, mastoid process, supra-orbital margin, supra-orbital ridge/glabella, and mental eminence) proposed by Walker (2008) were scored in this study for each individual, except for the mental eminence which could not be scored in eight U.S. white individuals because of significant alveolar resorption (Lewis & Garvin 2016). Inter- and intra-observer error rates were additionally calculated. Lewis and Garvin (2016) found that Observer A had an accuracy of 64.7% when scoring males and 85.7% when scoring females while Observer B had accuracy rates of 93.9% (males) and 92.6% (females). The intra-observer results demonstrated that there was considerable agreement for both observers for every trait, except the mental eminence. The inter-observer results indicated that the majority of the traits could be reliably scored, excluding the mental eminence (Lewis & Garvin 2016). The supraorbital ridge/glabella had the highest agreement and the mental eminence had the lowest (Lewis & Garvin 2016). Lewis and Garvin (2016) additionally found that prior experience had a statistical effect on trait scoring and sex classification. Simply put, the observer that had more experience with the method, as well as human variation produced greater accuracies with the estimations (Lewis & Garvin 2016). These

results produced by Lewis and Garvin (2016) had lower accuracy rates than those originally published by Walker (2008), where 99.5% of the repeated observations fell within the same range. Lewis and Garvin (2016) attributed this lower intra-observer agreement to both the experience level of the observers as well as potentially biased scoring when implementing the *gestalt* method.

Population Variation

The idea of population variation and its effect on sexual dimorphism in the skeleton underpins some of the existing methodologies behind population-specific models of sex estimation today. The hypothesis is that levels of morphological expression of sexually dimorphic features and traits vary between populations due to environmental, biogeographic, genetic, and nutritional differences (Klales 2020).

Generally, even across populations, male skulls tend to be more robust, larger in size, and have heavier muscle attachments when compared to females, but research suggests that population differences exist (Cunha & van Vark 1991; Eleveth & Tanner 1990; Franklin et al. 2005; Kemkes & Gobel 2006; Tallman 2019; Tallman & Go 2018; Ubelaker & DeGaglia 2017), but the extent to which discrete sexually dimorphic cranial traits vary across populations is still the subject of many studies (Garvin & Ruff 2012). Most methods employed currently by forensic anthropologists to produce biological profiles were developed in North America, the U.S. specifically, on African- and European-derived individuals as well as precontact Native Americans (Tallman 2019). Such methods are being applied on a large scale to specific and non-specific populations and it is not understood whether or not these methods can accurately estimate biological

profiles for combinations of modern-day populations (Tallman 2019). Additionally, the accuracy of visually-based sex estimation methods is predicated on the osteologist's prior experience and familiarity with the specific populations being studied (Ubelaker & DeGalgia 2017; Walker 2008). Both of these factors affect the reliability of the currently utilized methods.

Variation in sexually dimorphic skeletal morphology across populations is a byproduct of growth and developmental differences (Ubelaker & DeGalgia 2017). This variation is also potentially related to the genetic and environmental factors which affect methodologies applied to human skeletal remains to estimate sex (Ubelaker & DeGalgia 2017). Relethford (1994) suggests that the craniometric variation in modern humans is analogous to genetic variation and Strand Viðarsdóttir et al. (2002) establish that the ontogeny of osseous human facial features occurs early in life. Historically, researchers have used these theories to justify the utility of population-specific methodologies, but recent research indicates that early morphology is not a predictor of adult morphology because ontogenetic trajectories vary drastically due to environmental factors (Ubelaker & DeGalgia 2017). Franklin et al. (2005) evaluated the reliability of sexual dimorphism and discriminant function analysis models on a skeletal sample from South Africa. The study sample was comprised of indigenous individuals from the country. The results of the study indicated that sex could be accurately classified 77-80% of the time. A study from Kemkes and Göbel (2006) that assessed a previously documented method of assessment of the mastoid triangle examined on a collective sample of 297 skulls from Germany (50 females and 50 males) and Portugal (25 female and 72 males). The results

of this study indicated that despite detectable sex differences, the differences were not reliable enough for use in forensic investigations. Friedline et al. (2015) describe the facial complex as functioning as a series of stratified vertical levels that develop and form from a shared template, starting with the frontal lobes of the cerebrum and moving downward to the mandibular complex. This results in a descending cause-and-effect type of development and morphological integration of the neurocranium, basicranium, and facial cranium (Lieberman et al. 2008; Friedline et al. 2015). The skeletal structures of the craniofacial complex that are associated with growth of the brain, spine, and orbit tend to mature first (Humphrey 1998; Moss & Young 1960). These regions are followed by the mid-facial region, parts of the cranial base, and terminate with regions associated with the masticatory apparatus (Humphrey 1998; Moss & Young 1960). Because the skull is a complex and integrated structure that is morphologically affected by the interaction of multiple, partially independent components, the exact developmental relationship among the brain, cranial base, and face is still lacking, and the factors that influence the morphological variation of defining features are incompletely understood.

Ubelaker and DeGalgia (2017) suggest that population variation should be considered when estimating sex from skeletal remains and that population-specific data be utilized. While previous experience and knowledge of the vast range of population variation can assist osteologists in their assessments and research, this prior experience can also introduce systematic sex estimation biases (Walker 2008). For example, females in certain Californian Indian groups could potentially be misclassified as male if the

osteologist lacked a familiarity with population-specific reduced mastoid dimorphism (Walker 2008).

A study by Tallman (2019) specifically examined the differences in sexual dimorphism between Japanese and Thai individuals and the efficacy of the nonmetric-scoring systems when applied to these groups. As aforementioned, the methods employed today by osteologists were developed on non-Asian, historical populations, so it is unclear how well these standard methods estimate sex in such cases (Tallman 2019). The sample used in this study was comprised in total of 1,506 adult skeletons, aged 17 years and older, from different skeletal collections from the greater Tokyo region, the central and northern regions of Thailand, and the Philippines (Tallman 2019). The traditional sexually dimorphic traits of the skull were scored following Buikstra and Ubelaker (1994) and Walker (2008), which included the nuchal crest, mastoid process, supraorbital margin, supraorbital ridge/glabella, and mental eminence. The results of this study indicated that the standard sex estimation methods that were developed on African-derived, European-derived, and precontact Native Americans, do not effectively classify Asian individuals due to the variation in sexually dimorphic, nonmetric trait expression between Asian and non-Asian groups (Tallman 2019). Specifically, the nuchal crest, glabella, and mental eminence did not perform as accurately on the sample populations when compared to Walker (2008) and Garvin et al.'s (2014) samples (Tallman 2019). Similar to the population-specific reduced mastoid dimorphism from Walker's 2008 study, this is likely due to reduced cranial sexual dimorphism exhibited in the Japanese and Thai groups relative to the non-Asian groups (Tallman 2019). Ultimately, despite

research into quantifying population differences in regard to sex estimation, there is a lack of literature and research on largescale patterns of sexual dimorphism and their probable causes (Tallman 2019; Walker 2008). More recently, the limitations and exclusive nature of early sex estimation methodology from researchers like Thomas Dwight, Wilton Krogman, T.D. Stewart are being recognized as regionally specific, and though somewhat relevant to North America, are not representative of global variation (Klales 2020).

The Role of Ancestry in the Biological Profile

Over the years, the role of ancestry in the biological profile has evolved and shifted and its importance questioned (Bethard & DiGangi 2020). The sex estimation method that is used when developing the biological profile is often dependent on the estimated ancestry of a decedent, but recently, professionals in the field of forensic anthropology are questioning whether or not ancestry is a valuable component of the biological profile or simply an antiquated method of classification (Adams & Pilloud 2020; Bethard & Digangi 2020; Edgar 2020; Ross et al. 2020; Spradley & Jantz 2020).

In a letter to the editor, Bethard and DiGangi (2020) argued for an immediate cessation of estimation of ancestry from morphological traits of the cranium due to an incomplete understanding of the functional origin of the traits, as well as their problematic role in the debunked concept of biological race. Bethard and DiGangi (2020) suggest that the way in which ancestry estimation in forensic anthropological case reports is currently being applied could potentially be hindering identification efforts.

Within the field of forensic anthropology, estimates of "race", "ancestry", "population affinity", or "bioaffinity" have been considered necessary for the process of human identification, and when used in tandem with other parameters such as age-at-death and sex, are estimates that provide inclusionary and exclusionary parameters with which to apply to a population of missing persons or unidentified remains (Bethard & DiGangi 2020). With the possibility of misidentification or missed identifications, as well as the social implications of socially constructed "race" categories, it is imperative that the forensic anthropological discipline critically evaluate traits that are not fully understood (Bethard & DiGangi 2020). Bethard and DiGangi (2020) emphasize that morphometrics are representative of a bygone era where the anthropologists at the forefront of the field believed that *Homo sapiens* could be separated into discrete groups.

Researchers maintain that estimating ancestry using cranial morphometrics and metrics is rooted in the way that human variation is reflected in the skeleton and ultimately linked to heritability (Carson 2006a, 2006b). Additionally, forensic anthropologists have historically functioned under the explicit notion that ancestry is an essential part of the biological profile and that, despite skeletal ancestry and skin color not being the same thing, the estimations and information from their reports are directly compared with lists of missing people (Bethard & DiGangi 2020). Bethard and DiGangi (2020) postulate that this seemingly innocuous parameter has the potential to negatively impact the positive identification rate of marginalized populations, such as African, Asian, Indigenous, or Latin American people, due, in part, to the "Missing White Woman Syndrome" (Sommers 2016; Moss 2018). The Missing White Woman Syndrome

(alternatively, Missing White Girl Syndrome) manifests as an overabundance of attention and exposure from mainstream media outlets to missing person's cases of white women and a correlating lack of exposure for cases of missing people of color (Moss 2018). Sommers (2016) suggests that victim media representation is affected by a hierarchy. More specifically that non-white victims are disproportionately underrepresented in the media coverage when compared to their white counterparts, and that white women and girls, in particular, experience higher rates of publication (Sommers 2016). Moss (2018) implies that this disparity will continue to intensify as technology evolves into the primary method for accessing news and current events.

As a result of this syndrome and the potential for racial bias on behalf of the investigators, there is a risk that unidentified decedents may be subjected to inadequate investigative efforts following an ancestry estimation of anything other than European-descended (Bethard & DiGangi 2020).

The morphological traits of the skeleton that have been studied by biological anthropologists and population geneticists to elucidate patterns of human biological variation are not the traits examined and used by Forensic anthropologists to estimate ancestry (Bethard & DiGangi 2020). This suggests that research that includes these nonmetric traits has not considered evolutionary theory or ecogeographic variation which are foundational theories for biological anthropology (Bethard & DiGangi 2020). If ancestry (or race, population affinity, bioaffinity) is phased out of the biological profile, then sex estimation methods will need to be revisited and adapted for non-specific populations.

Age-Related Effects on Sex Estimation

As humans age, their skeletons undergo degenerative processes that change the morphology of the bone. Sex estimation, which mostly relies on nonmetric and morphological assessment, can potentially be affected by the aging process (Klales 2016; Lovell 1989; Sutherland & Suchey 1991; Vollner et al. 2014). Some research shows that the pelvis is at a greater risk than the cranium of being affected by morphological change as a result of aging (Lovell, 1989; Sutherland & Suchey, 1991), but other research has observed no significant age-related disturbances in sex estimation using the pelvis (Klales, 2016; Sutherland & Suchey, 1991; Vollner et al. 2014). Lovell (1989) suggests that the identification of the ventral arc becomes more difficult in older individuals due to degenerative age-related changes to the pubis which may impact expression of morphological traits. On the other hand, Sutherland and Suchey (1991) observed higher frequencies of ventral arc prevalence in females that were up to 90 years old, contradicting Lovell's (1989) findings and ultimately demonstrate that the ventral arc is useful in females of all ages as an estimator of sex. Additionally, Vollner et al. (2009) found no correlation between an individual's age and trait expression score for the three traits most commonly used in estimating sex: the medial aspect of the ischio-pubic ramus, ventral arc, and subpubic contour. Kales (2016) observed that although it is never absolutely certain how the different age distributions might affect the differences observed in trait score frequencies over time in the pelvis, previous research (Sutherland & Suchey, 1991; Vollner et al., 2009) has indicated that the different age distributions do

not have a significant impact on the results obtained in studying the secular trends in morphological pelvic traits.

In Walker's (2008) previously discussed study of nonmetric cranial traits, the analysis of the scoring system also revealed variable distributions of trait scores in the study populations of modern U.S. Black, U.S. white, English, and ancient Native Americans. Walker (2008) found a significant effect of age on cranial trait distribution that corroborated previous research on the subject (Walker 1995; Walker 2005), specifically that supraorbital robusticity increased in males over thirty years old and in females over forty-five years of age. Conversely, Garvin et al. (2014) evaluated the same morphological cranial trait scores from Walker's (2008) study, as well as postcranial measurements, from four diverse population groups to re-evaluate the potential effects of population, age, and body size on sexual dimorphism of the skeleton. The study sample included a total number of 499 individuals from US white, US Black, medieval Nubia, and Arikara Native American populations. Ultimately, Garvin et al. (2014) found that neither age nor body size played a significant role in trait expression, contradicting Walker's (2008) findings and indicating that these variables did not need to be factored into methods of sex estimation.

Tallman (2019) assessed the nonmetric sexually dimorphic traits of the cranium in conjunction with sex estimation in east and southeast Asian individuals. Analyses showed a slight positive correlation between the nuchal crest scores and age for the Japanese and Thai males (Tallman, 2019). This suggested that as age increased, the nuchal crest was becoming increasingly more robust. Additionally, for females, there was a slight negative

correlation between the score for the mental eminence and age that indicated an increase in gracilization with age (Tallman, 2019). Overall though, age did not significantly affect the cranial nonmetric traits in the Japanese and Thai males and females (Tallman, 2019).

Computed Tomography (CT) Validation Studies

The use of CT scans has been shown to be an acceptable alternative method of data collection to traditional analysis with dry bone (Decker et al. 2011; Hughes et al. 2013; Richard et al. 2014; Ross 2004; Stull et al. 2014) and this type of advancement in technology has allowed researchers to attempt identification methods with CT scans of skeletons, without the removal of soft tissue (Stull et al. 2014). The advancement of computed tomography (CT) scans is providing researchers with examination capabilities beyond in-person observation (Decker et al., 2011). This type of capability can provide universal accessibility in times where working hands-on with skeletons is not possible due to fragility of the remains, other safety reasons, or in cases where skeletons are not available for the construction of estimation models, anthropologists can access CT databases for suitable data (Stull et al. 2014). A study by Decker et al. (2011), examined 100 individuals with known demographics and complete pelves in order to test the accuracy of virtual determination of sex (nonmetric and metric) using land-marking and measurement, metric evaluation, and a technique derived from Phenice (1969). The tests resulted in an overall accuracy rate of 100%, demonstrating that it is possible to estimate sex accurately from CT scans (Decker et al., 2011). This is especially important as it shows that medical imaging and modeling enable remote analyses without assuming chain of custody of the evidence, assuring that data is securely transferred to other

investigators and experts (Decker et al., 2011). Additionally, this type of data can replace traditional autopsy and can be re-examined even after the original, physical remains have been buried, cremated, or deteriorated (Decker et al. 2011).

Richard et al. (2014) examined the accuracy of standard craniometrics measurements using multiple data formats. Richard et al. (2014) tested the reliability of 26 standard craniometrics measurements most often employed in forensic casework across several imaging technologies. The study sample was comprised of five crania of known individuals of which data was collected via computed tomography (CT) scans and three-dimensional (3D) laser scans. The measurements gathered from each technology-type were compared to measurements collected using traditional osteometric methods of measuring. The results of this study indicated that the majority of measurements were not significantly different across the technology formats. Measurements taken in a supero-inferior direction such as basion-bregma height and orbital height were subject to greater deviations from direct measurement of the crania. Despite relative consistency across technology formats, Richard et al. (2014) cautioned researchers against using medical scans in place of measurements taken on an actual skull unless necessary.

Stull et al. (2014) specifically assessed the accuracy and reliability of measurements obtained from 3D-volume rendered CT images. The study sample was comprised of a full-body CT scan of a deceased individual, performed with a General Electric (GE) Light Speed RT-16 multi-detector scanner before they were autopsied at the Office of the Chief Medical Examiner (OCME) for the State of Maryland. Following the full-body CT scans, the decedent was processed to extract certain skeletal elements (cranium,

mandible, left clavicle, right and left scapula, humeri, ulnae, radii, femora, tibiae, and fibulae), and the soft tissue was then removed from these skeletal elements which were then processed again to dry and when dry the elements were rescanned. In total, three sources of measurements were obtained including (1) dry skeletal elements, (2) CT images with soft tissue before autopsy, (3) and CT images of the dry skeletal elements after soft tissue removal. Seven additional crania were processed in the same way in order to assess reliability and repeated measurements. The primary concern of this study was whether or not the measurements collected on the CT images were as accurate as the data collected on the dry skeletal material. The average percent difference for all measurement types combined was 1.4% for the dry-dryCT comparison, 1.5% for the dry-CT comparison, and 2.9% for the dryCT-CT comparison. These results indicated that CT assessment is a suitable replacement for dry bone examination. Any reduction in accuracy was attributed either to the differential Hounsfield units of air and soft tissue or susceptibility of the CT image to certain artifacts like partial volume effects (PVE) (Stull et al. 2014).

In conclusion, recent developments in the field of forensic anthropology have brought to light the problematic nature of currently used ancestry estimation methods in that the conventional models used have generally all been developed and tested on specific populations of U.S. Black, white, and occasionally Native American skeletons (Tise et al. 2103). Additionally, the ancestry methods have not taken into consideration the evolutionary mechanisms behind biogeographic differences, or the influence that secular

change, intersectionality, and biocultural processes on skeletal variability across populations (Albanese and Saunders 2006; Bethard and DiGangi 2020; Blakey 2021; Carson 2006a; DiGangi and Bethard 2021). By association, the conventional practice of applying sex estimation models fundamentally reliant on these flawed ancestry estimation models is not an appropriate or accurate method of analysis. In short, the application of incompatible population-specific models and uninformed ancestry estimation should be left in the past with the typological methodologies that are intrinsically linked to it.

In cases where ancestry is unknown and in light of the ongoing debate regarding the role of ancestry in the biological profile (Bethard & DiGangi 2020), a population-inclusive nonmetric and metric model for estimating sex that does not require ancestry to be estimated could be a vital piece of the forensic anthropologists' toolkit.

CHAPTER THREE: METHODS

The objective of this research was to compare the performance of population-inclusive and population-specific sex estimation models in order to assess the necessity of ancestry estimation in the development of the biological profile. The expectation for this research followed that the rate of individuals correctly classified using the population-inclusive model would not be significantly different than the population-specific models developed with this research and those traditionally and currently used by forensic anthropologists. Evidence in support of this hypothesis would suggest that a population-inclusive model could be used in lieu of population-specific models when developing the biological profile, but more specifically, estimating sex. This would allow forensic anthropologists to omit ancestry or population affinity from the biological profile. 3D-CT images of the skull from a combination of five population affiliation groups (including African American, Asian American, European American, Latin American, and Native American) comprised the study sample used to develop the population-inclusive models to address this research question.

Ultimately, this analysis aimed to evaluate the overall accuracy of the traditional population-specific models and the margin of error produced when applying population-specific models to groups different from the reference population and to compare the rate of correct classification between population-inclusive and specific models. Furthermore, the degree of similarity between accuracy rates for population-inclusive and specific sex estimation models has implications for the necessity of ancestry estimation in the development of the biological profile.

Study Sample

This study incorporated anonymized full-body CT scans of decedents that were routed to the Office of the Chief Medical Investigator (OCMI) for Albuquerque, NM for autopsy between 2010 and 2017 (Daneshvari Berry & Edgar 2021). The images were originally collected to evaluate the utility of supplementing traditionally autopsies with CT scans and eventually compiled into the New Mexico Decedent Image Database (NMDID) for further research purposes. Because the decedent metadata was completely anonymized, the protocol for this study was granted an exemption through Boston University Institutional Review Board (H-40441).

The study sample is comprised of 431 individual 3D-CT images of the skull – originally 494 and later reduced due to downloading errors and incompatible volume-rendering that affected approximately 13% of the original sample. The CT images were first collected from the NMDID, with each individual represented by an average of 2,000 axial image slices of 1 mm, with 0.5 mm overlap and a 512 x 512 matrix. Each individual was also accompanied by scout images similar to whole body radiographs. These were examined to briefly assess the condition of the remains before being officially selected for the sample pool.

Individual decedents were selected based on sex, age, and population affiliation. Population affiliation/affinity was established/estimated based on the self-identified or next-of-kin-identified race, ancestry, and/or ethnicity recorded in the NMDID. The exclusionary criteria included. (1) causes of death that would impede data collection of the skull (e.g., blunt force trauma and gunshot wounds to head and neck, thermal

injuries); (2) ages not contained within the range of 18-90 years; and (3) individuals who did not identify with the male or female sex assignment at birth. These exclusionary criteria were applied when possible, but were limited by the set study cohorts as well as the diversity of the NMDID, which is disproportionately comprised of males.

The criteria for inclusion encompassed female and male (sex assigned at birth) individuals between 18 and 90 years from one of the five pre-determined population affinities. In this study sample, there are 242 males and 189 females. Four age cohorts were established to ensure a similar distribution across both sex categories (Figure 3.1) as well as across population affinities (Figure 3.2). The age cohorts are 18-30 years (n=109), 31-50 years (n=114), 51-70 (168), and 71-90 years (n=40) (Table 3.1). The five population affinity groups created for this research were constructed from social race categories of the U.S. census as a proxy for population affiliation, as well as the "Physical Characteristics" subcategories of "race" and "ethnicity" of the NMDID and are relatively equally distributed across all five categories (Figure 3.3). These bureaucratic groups do not capture the entire range of human diversity but are representative of the five major demographics listed on U.S. government data collection forms. The final groups created for this study include African American (original NMDID variable: Black or African American (race); m=40, f=43), Asian American (original NMDID variable: Chinese, Filipino, Japanese, Korean, Vietnamese, and Other Asian (race); m=69, f=20), European American (original NMDID variable: white (race); m=43, f=41), Latin American (original NMDID variable: Hispanic or Latino (ethnicity); m=44, f=43), and Native American (original NMDID variable (race): not broken down by tribe, m=46, f=42). One

generalized Asian-derived population affiliation was constructed from the separate "race" categories of Chinese, Filipino, Japanese, Korean, Vietnamese, and Other Asian because these categories did not have a large enough representation within the database compared to the other population affiliations. When selecting individuals for all population groups excluding Latin American, ethnicity was additionally selected to be "Not Hispanic, Latino, or Middle Eastern" to ensure no cross listing between the categories. When the Latin American group was constructed the Hispanic and Latino ethnicity was selected.

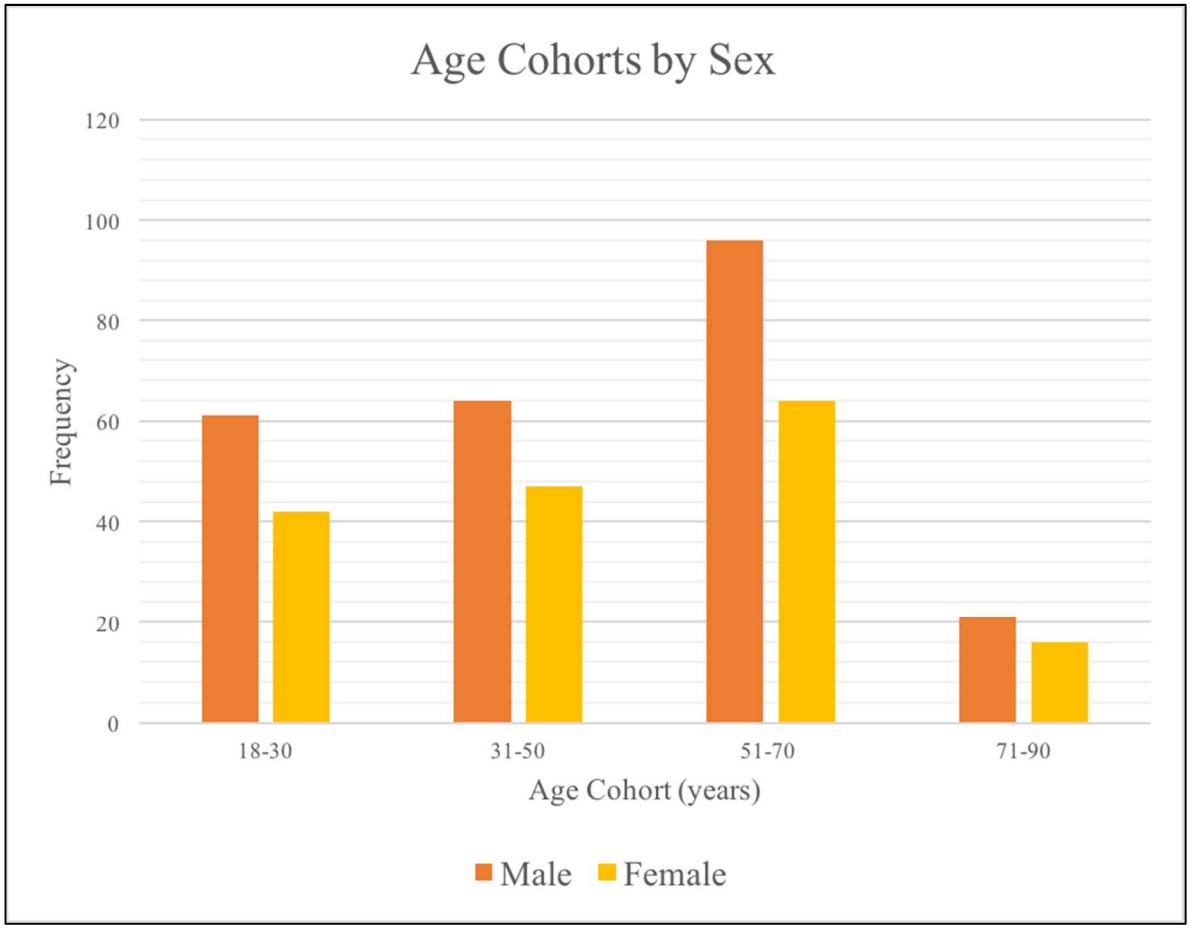

Figure 3.1. Age Cohorts by Sex in Study Sample.

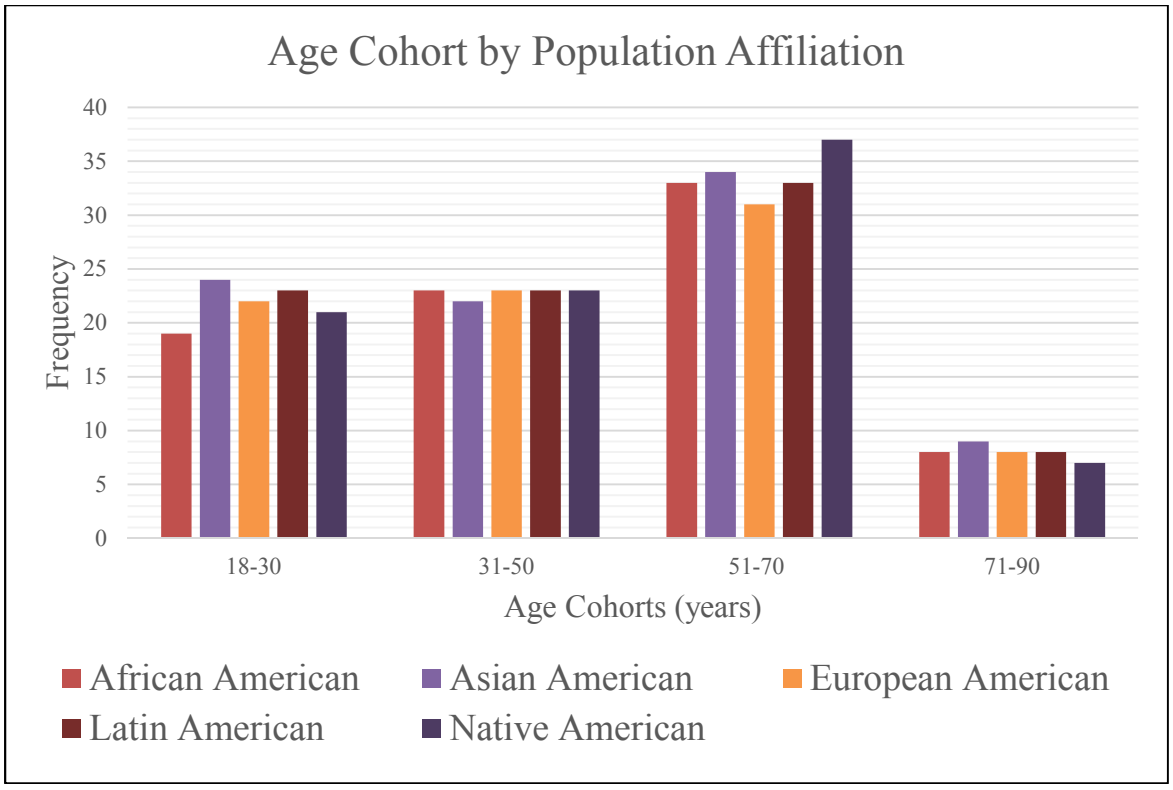

Figure 3.2. Age Cohorts by Population Affiliation in Study Sample.

Table 3.1. Study Sample Broken Down by Age Cohort, Population Affiliation and Sex.

Age (years)	*Population Affiliation*	*Male*	*Female*
18-30	African American	10	9
	Asian American	18	6
	European American	10	12
	Latin American	12	11
	Native American	11	10
31-50	African American	11	12
	Asian American	19	3
	European American	12	11
	Latin American	11	12
	Native American	11	12
51-70	African American	15	18
	Asian American	26	8
	European American	18	13
	Latin American	17	16
	Native American	20	17
71-90	African American	4	4
	Asian American	6	3
	European American	3	5
	Latin American	4	4
Total Sample Size		238	186

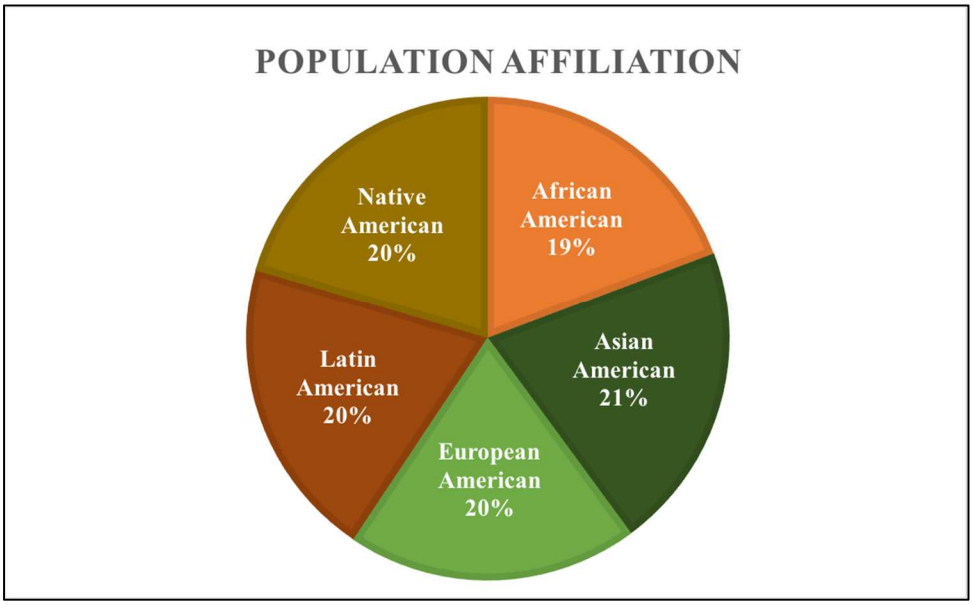

Figure 3.3. Demographics of Study Sample by Population Affinity.

Sample Preparation

The preparation of the samples consisted of a multi-step process. The CT-image files for each individual were downloaded and visualized using the 64-bit version of *OsiriX MD* Imaging Software (v.11.0.4) (Pixemo Sarl 2016) on an iMac™ computer. Using this software, 3D images of the osseous structures were rendered using the 3D volume rendering function with a focus on bone. This image was then converted to a Meshmix file using Meshmixer™ in order to process the original image. The editing process consisted of removing life-saving equipment (e.g., defibrillator pads, tubes, wiring, clamps, etc.) as well as personal artifacts (e.g., eyeglasses, jewelry, hair pins, buttons, etc.) and other artifacts (i.e., partial volume effects). The aim of this editing process was to create a sample of 3D images that consisted of the isolated skull, as well as to ensure accessibility of the features that were scored and surfaces where points were

placed for metric measurement (e.g., removal of first cervical vertebrae for access to foramen magnum as well as basion), and remove any potential for biases (e.g., removal of jewelry, hair pins, etc.) (Figure 3.4). This process took approximately 10-20 minutes per individual.

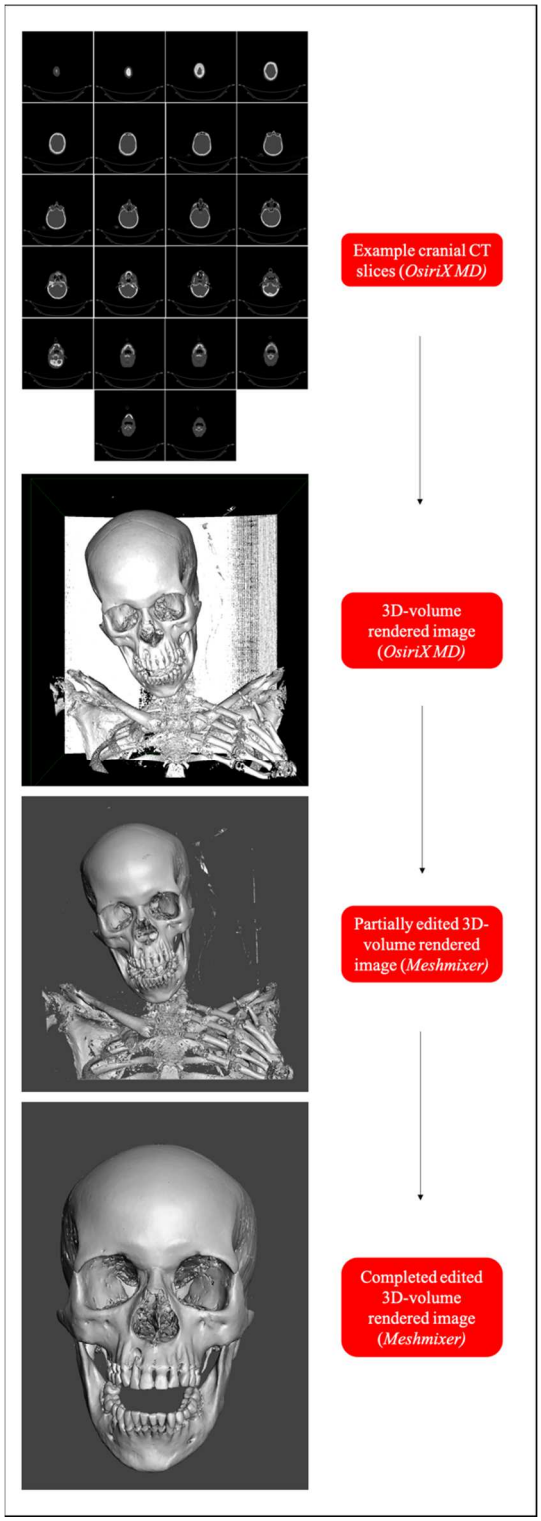

Figure 3.4. Sample Preparation Example Including the Progression of an Individual from CT-Image Slices, to 3D Volume-Rendered CT Image, to Completed Edited Mesh File.

Sample Data Collection

Nonmetric Scoring of the Skull

Nonmetric traits were scored following Buikstra and Ubelaker (1994) and Walker (2008) (Figure 3.5) and included the supraorbital ridge/glabella, supraorbital margin, mastoid process, nuchal crest, and mental eminence. The 3D-CT skulls were visually examined from the same angle each time and compared with the graphics (Figure 3.5) and descriptions (Table 3.2) from Buikstra and Ubelaker (1994). The traits were each assigned an ordinal score on a scale from 1 to 5 (gracile to robust, respectively). Traits were only scored if they were complete or mostly complete in the case of the mastoid, if at least one was represented, and if the surrounding structures were intact enough to provide relative comparison. The mental eminence was not scored if there was significant alveolar resorption.

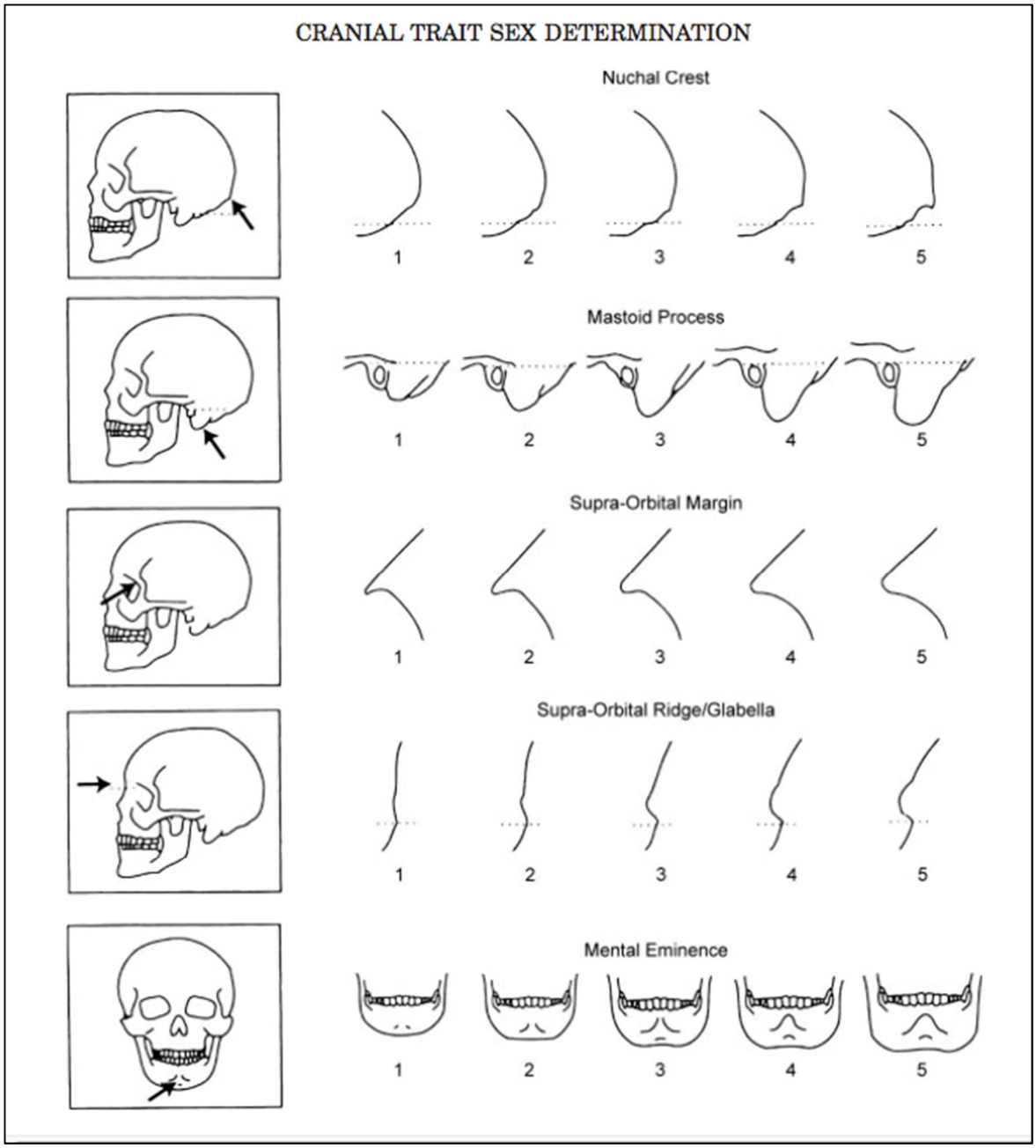

Figure 3.5. Nonmetric Traits (Buikstra & Ubelaker 1994:20 and Walker 2008:41).

Table 3.2. Descriptions of Nonmetric Trait Expression Adapted from Buikstra and Ubelaker (1994:19-20).

Nonmetric Trait	Description
Nuchal Crest	View the lateral profile of the occipital and compare it with the diagrams. Feel the surface of the occipital with your hand and note any surface rugosity, ignoring the contour of the underlying bone. Focus upon the rugosity attendant to the attachment of nuchal musculature. In the case of minimal expression (Score = "1"), the external surface of the occipital is smooth with no bony projections visible when the lateral profile is viewed. Maximal expression (Score = "5") defines a massive nuchal crest that project a considerable distance from the bone and forms a well-defined bony ledge or hook
Mastoid Process	Score this feature by comparing its size with that of surrounding structures such as the external auditory meatus and zygomatic process of the temporal bone. Mastoid processes vary considerably in their proportions. The most important variable to consider in scoring this trait is the volume of the mastoid process, not its length. Minimal expression ("1") is a very small mastoid process that projects only a small distance below the inferior margins of the external auditory meatus and the digastric groove. A massive mastoid process with lengths and widths several times that of the external auditory meatus should be scored a "5".
Supraorbital Margin	Begin by holding your finger against the margin of the orbit at the lateral aspect of the supraorbital foramen. Then hold the edge of the orbit between your fingers to determine its thickness. Look at each of the diagrams to determine which it seems to match most closely. In an example of minimal expression ("1"), the border should feel extremely sharp, like the edge of a slightly dulled knife. A thick, rounded margin with a curvature approximating a pencil should be scored as a "5".
Supra-Orbital Ridge/Glabella	Viewing the cranium from the side, compare the profile of the supraorbital region with the diagram. In a minimal prominence of glabella/supraorbital rides ("1") the contour of the frontal is smooth, with little or no projection at the midline. Maximal expression involves a massive glabellar prominence, forming a rounded, loaf-shaped projection that is frequently associated with well-developed supraorbital ridges.
Mental Eminence	Hold the mandible between the thumbs and index fingers with thumbs on either side of the mental eminence. Move the thumbs medially until they delimit the lateral borders of the mental eminence. In example of minimal expression ("1"), there is little or no projection of the mental eminence above the surrounding bone. By contrast, a massive mental eminence that, occupies most of the anterior portion of the mandible is scored as "5".

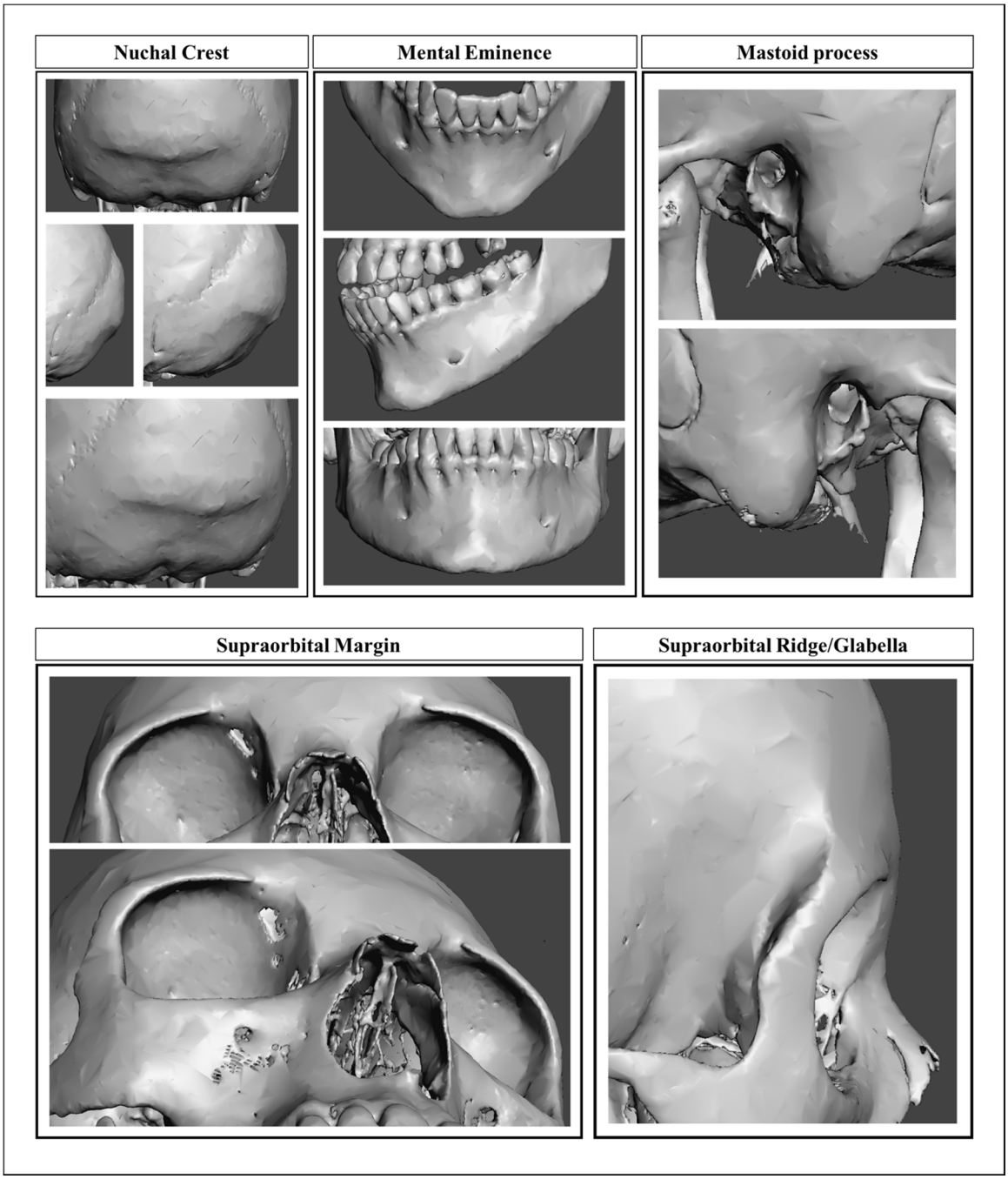

Figure 3.6. Example of Aspects used to view the Five Morphological Traits.

Metric Measurements of the Skull

Metric measurements of the skull followed the methods Buikstra and Ubelaker (1994) and incorporated 18 standard points of measurement of the cranium and the mandible (Figures 3.7-3.13) from Spradley and Jantz (2011). The cranial measurements included glabella occipital length, bizygomatic breadth, basion-bregma height, basion-nasion length, biauricular breadth, upper facial height, minimum frontal breadth, nasal height, orbital height, frontal chord, parietal chord, occipital chord, foramen magnum breadth, and mastoid height. The mandibular measurements included bigonial breadth, bicondylar breadth, maximum ramus height, and mandibular length. Using the Meshmixer ™ software, metric measurements were collected using the 'Units/Dimensions' function where the points were placed on the appropriate cranial landmarks and the *Real Length* in millimeters (mm) was recorded and rounded to the nearest thousandths. For maximum and minimum measurements, the points were dragged to capture the highest or lowest number produced.

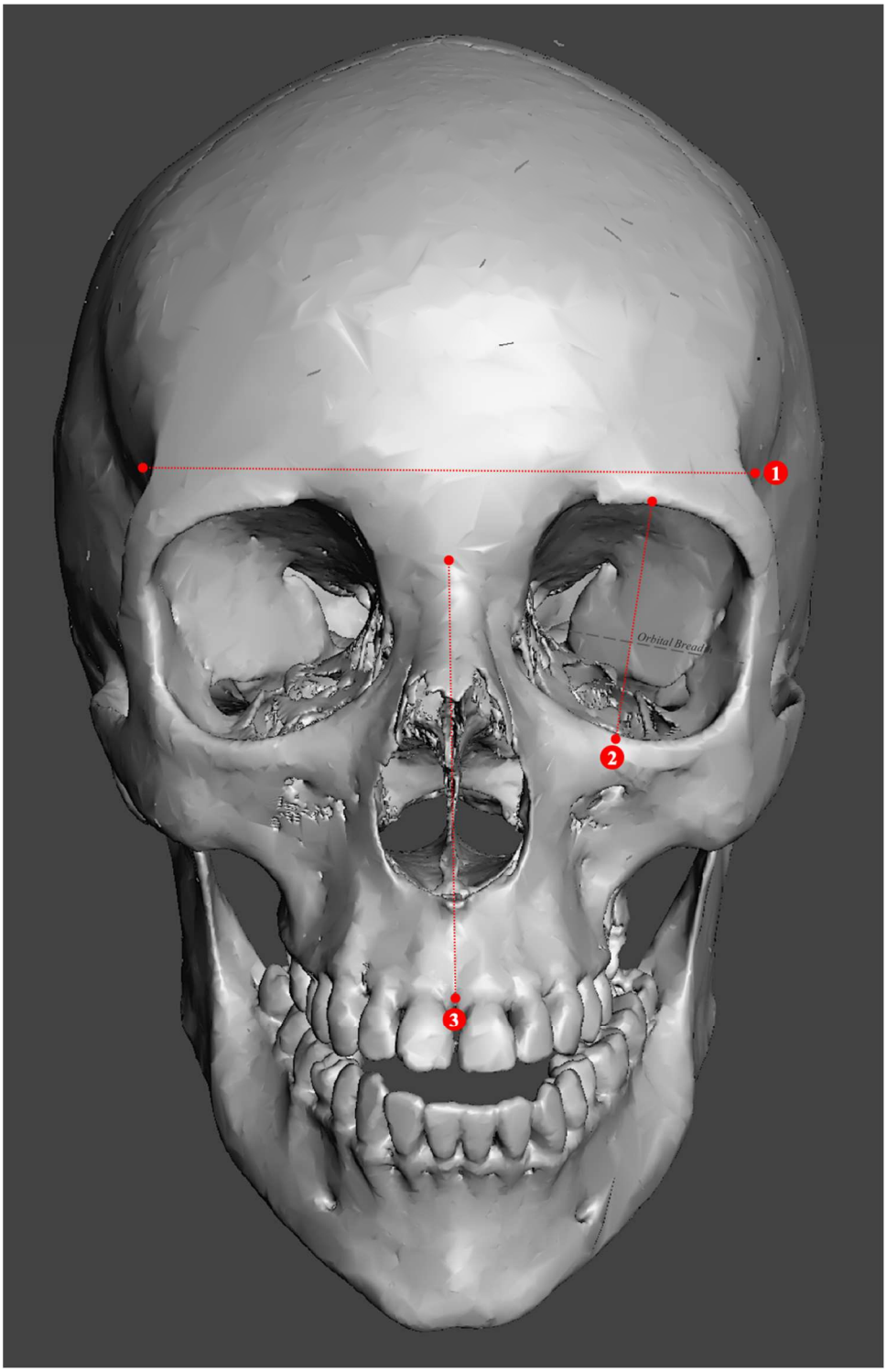

Figure 3.7. Example of Metric Measurements from Anterior Perspective, Including (1) Minimum Frontal Breadth, (2) Orbital Height, and (3) Upper Facial Height.

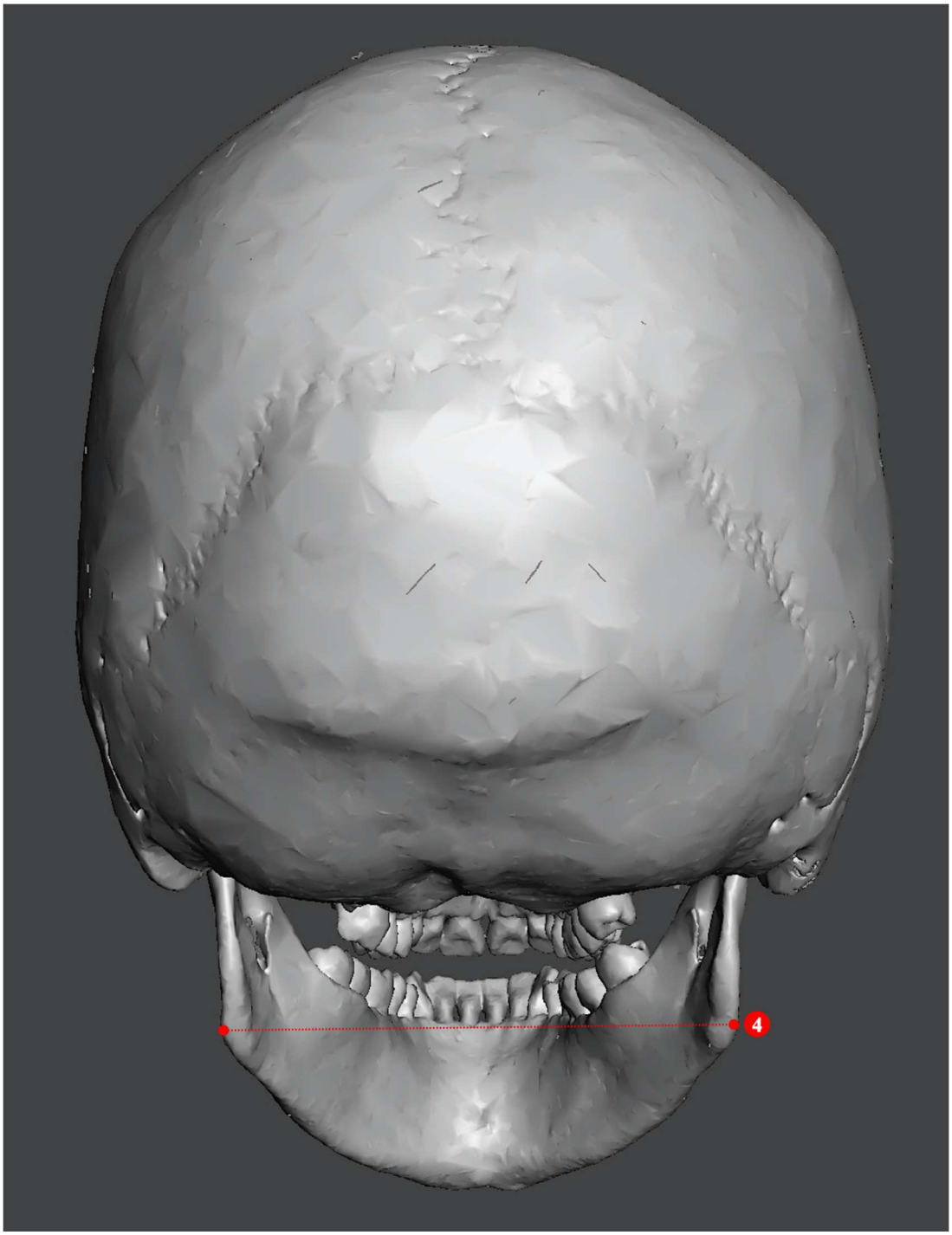

Figure 3.8. Example of Metric Measurements from Posterior Perspective, Including the (4) Bigonial Breadth.

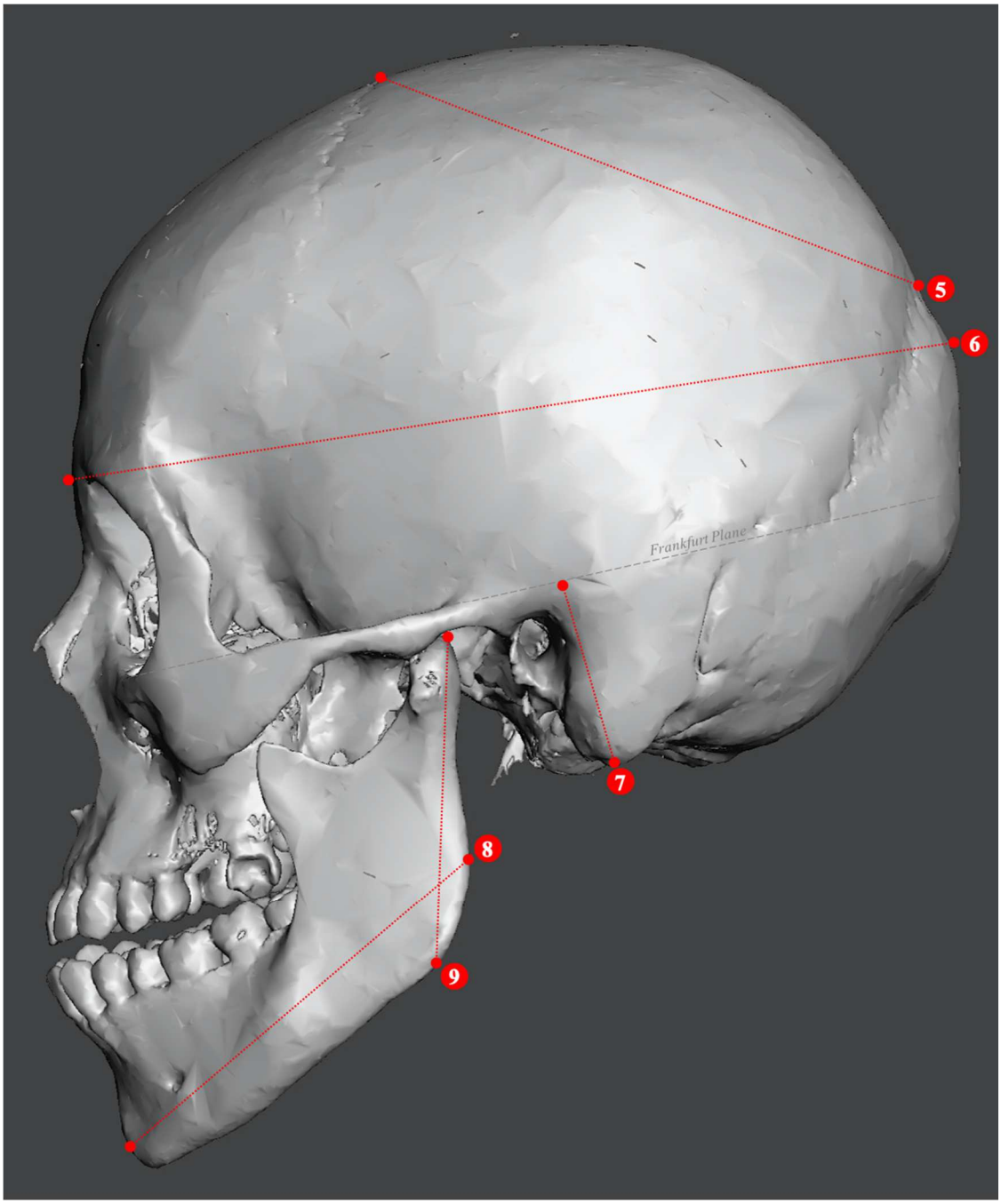

Figure 3.9. Example of Metric Measurements from Lateral Perspective, Including (5) Parietal Chord, (6) Occipital Chord, (7) Mastoid Length, (8) Mandibular Length, and (9) Maximum Ramus Height.

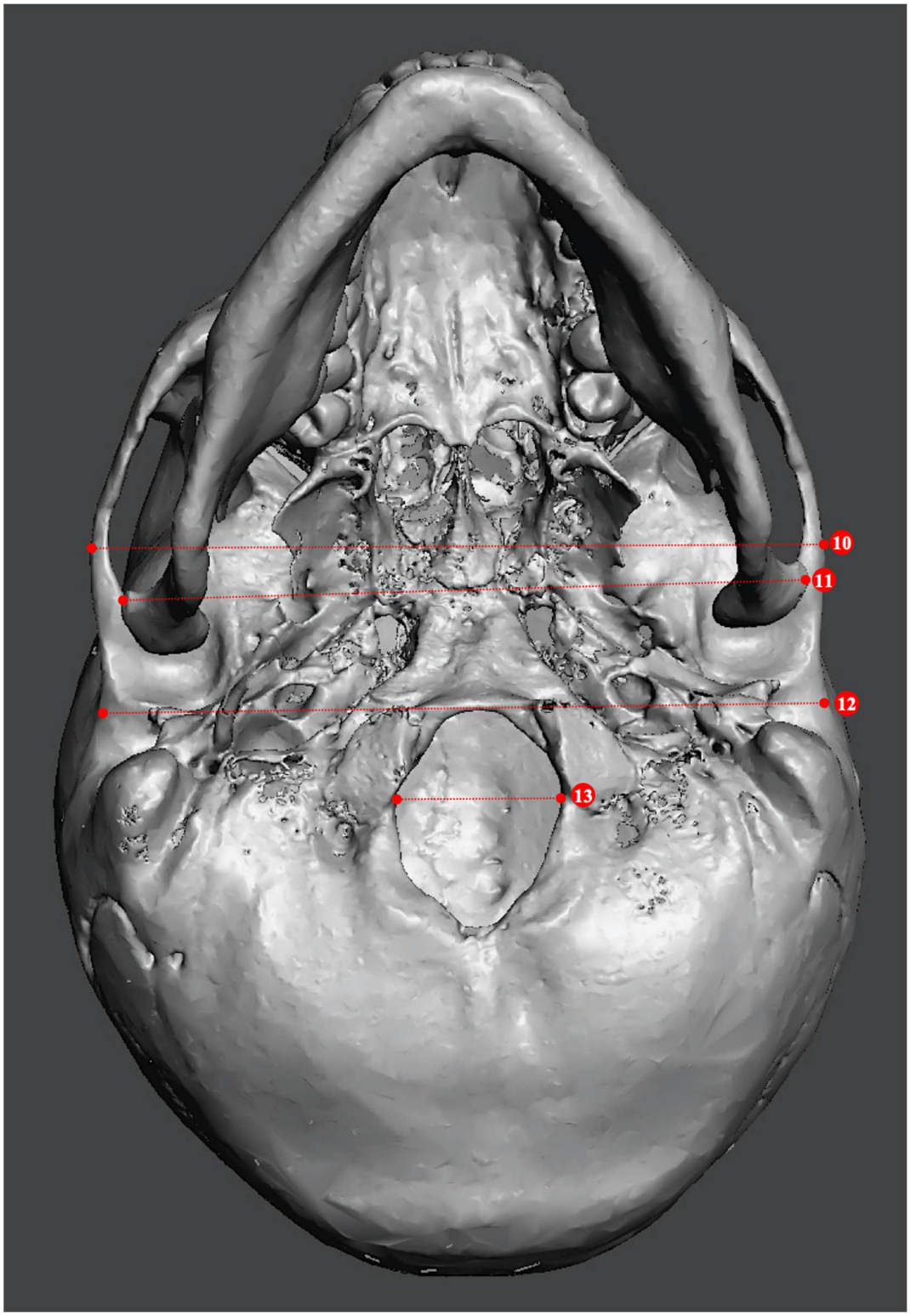

Figure 3.10. Example of Metric Measurements from Inferior Perspective, Including (10) Bizygomatic Breadth, (11) Bicondylar Breadth, (12) Biauricular Breadth, and (13) Foramen Magnum Breadth.

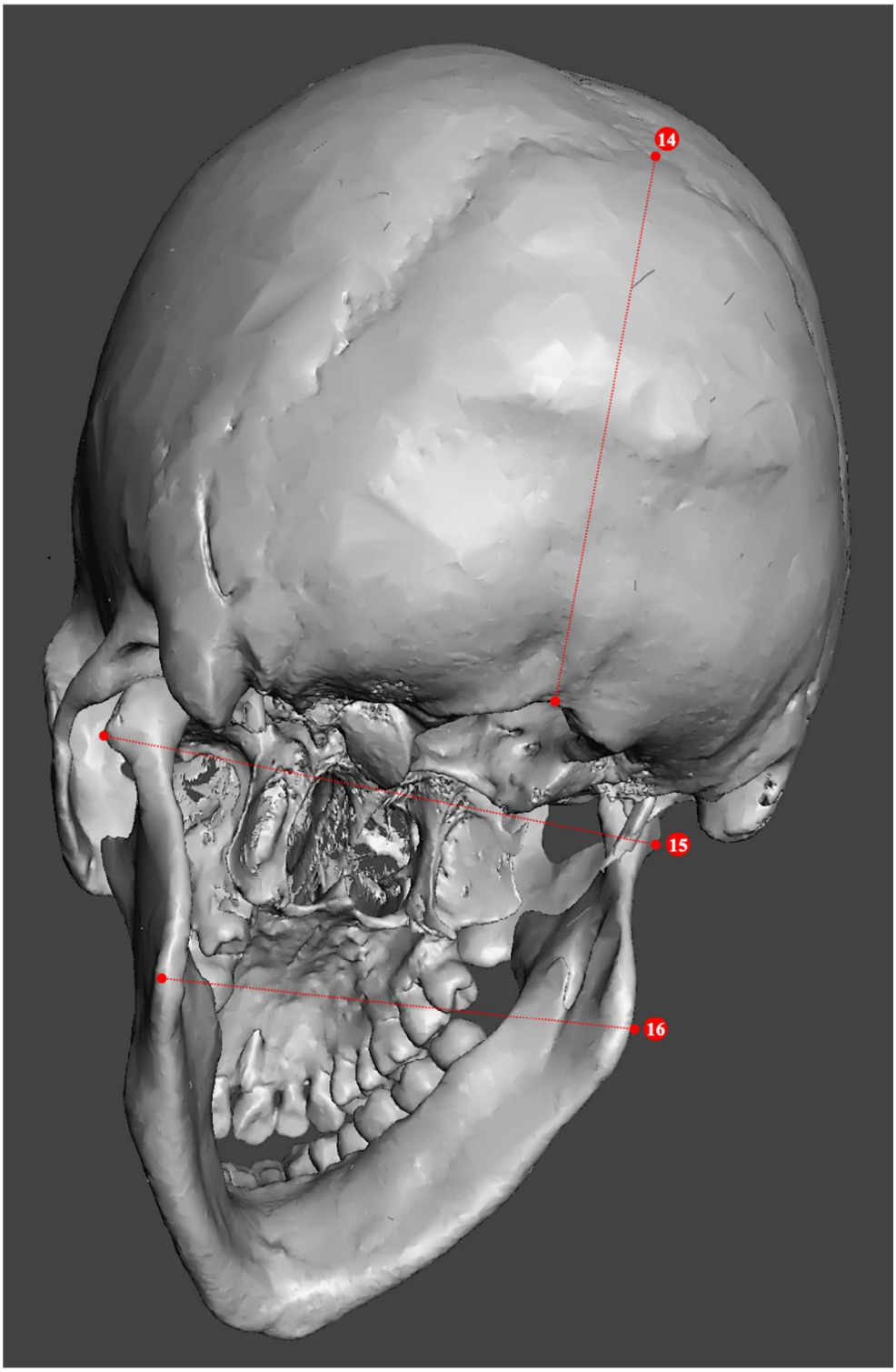

Figure 3.11. Example of Metric Measurements from Inferior Perspective, Including (14) Occipital Chord, (15) Bicondylar Breadth, and (16) Bigonial Breadth.

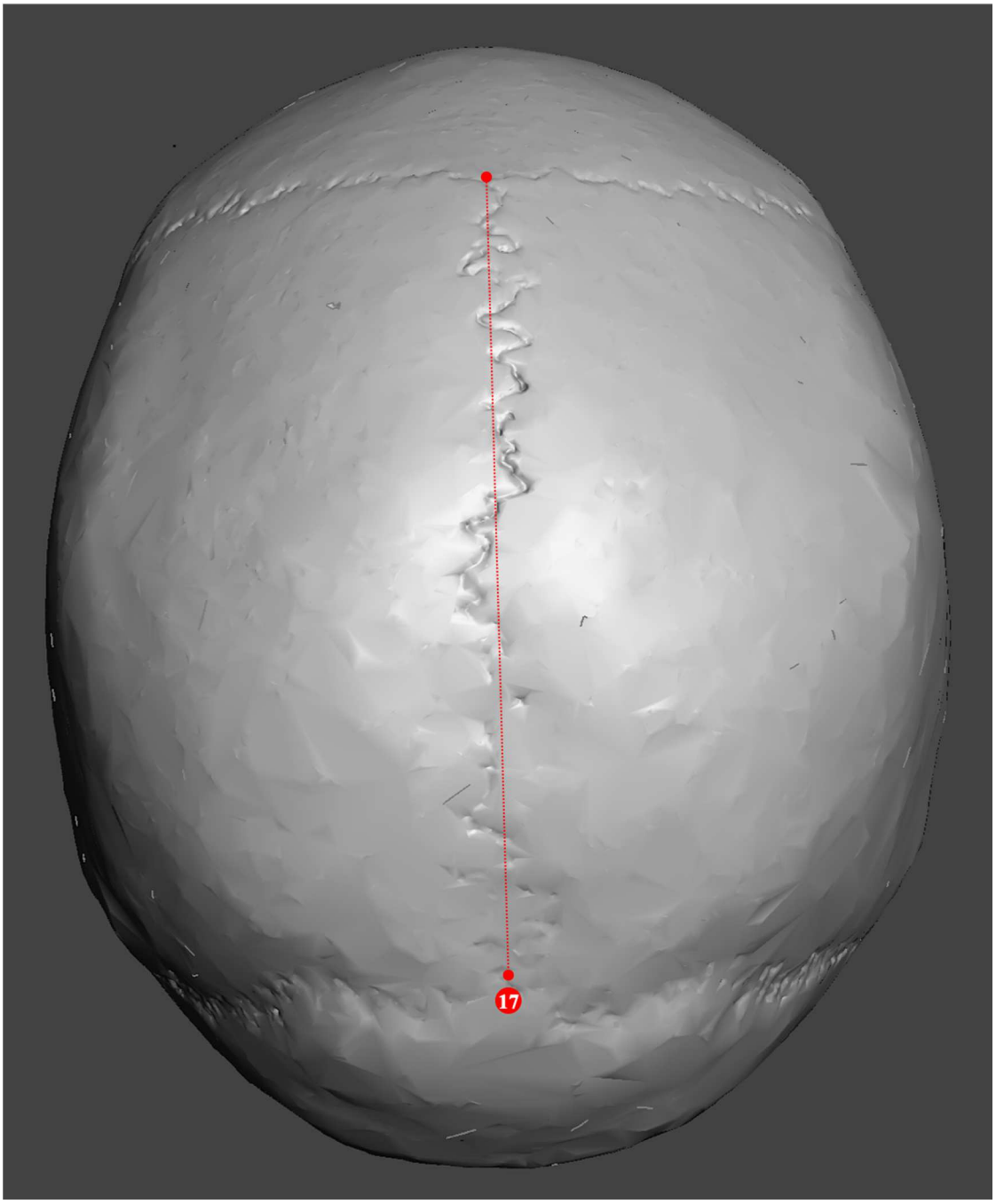

Figure 3.12. Example of Metric Measurements from Superior Perspective, Including (17) Parietal Chord.

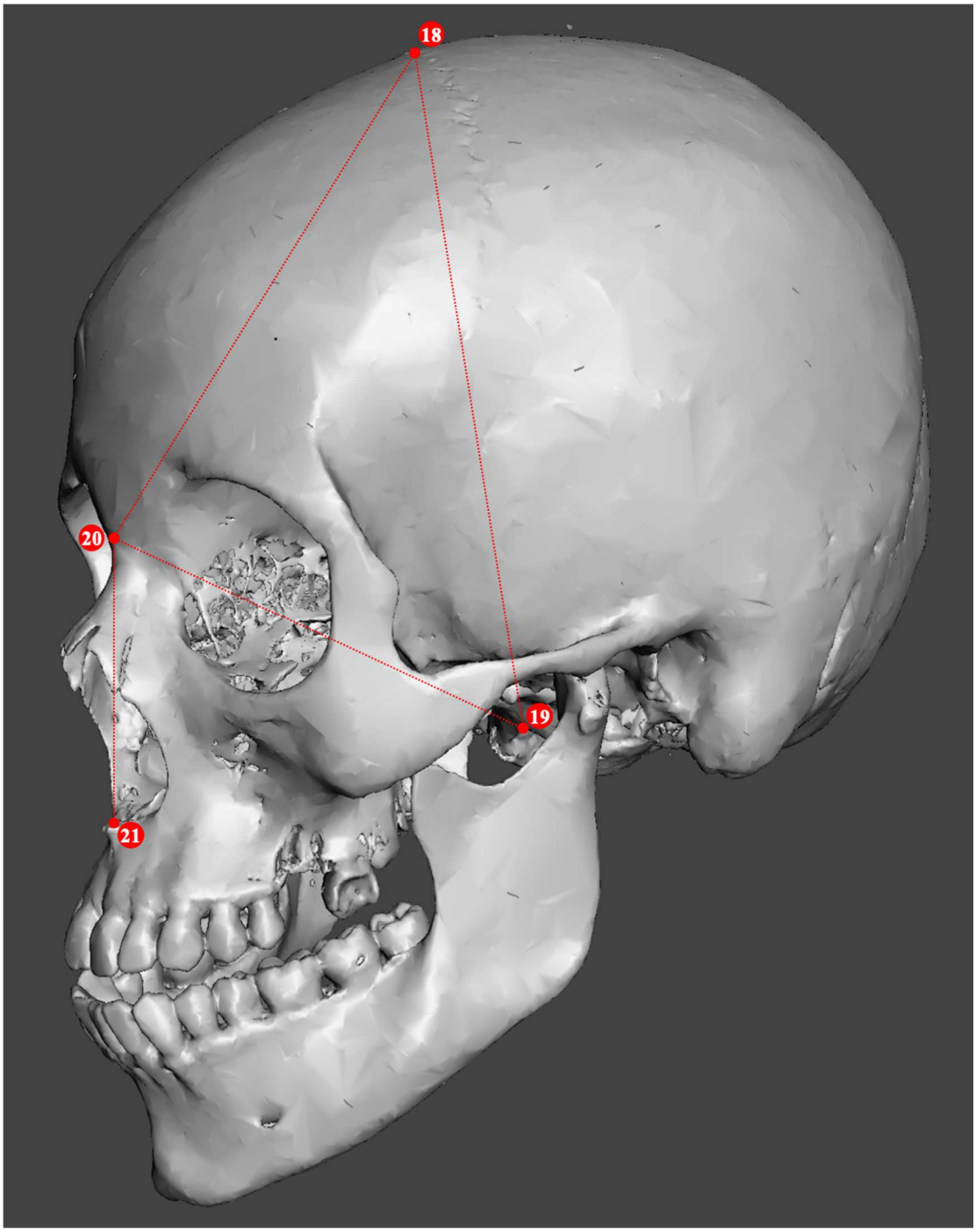

Figure 3.13. Example of Metric Measurements from Lateral Perspective, Including (18) Basion-Bregma Height, (19) Cranial Base Length, (20) Frontal Chord, and (21) Nasal Height.

Statistical Analyses

All statistical analyses were run with IBM SPSS (version 27). Binary logistic regressions (BLR) and discriminant function analyses (DFA) were employed in order to develop population-specific and population-inclusive models. The BLR analysis was employed to analyze the nonmetric data as well as a combined nonmetric and metric data model. A DFA was employed to assess and the metric data and create a metric model. Two-way cross tabulation tables were created in SPSS using Fisher's exact test based on the Chi-square statistic in order to evaluate whether significant differences existed between the population-inclusive and population-specific models (Hefner 2009).

Binary Logistic Regressions for Nonmetric and Combined Models

A BLR was used to analyze the nonmetric traits because BLR analyses does not require normal distributions (avoids problem of violating assumptions of linearity) and is well suited to ordinal scoring methods (Fields 2009:267; Hefner 2009; Walker 2008). The Forward: Wald stepwise selection method was used so that variables would be removed or tested based on significance. SPSS set the cut value as 0.5 so that individuals with probabilities less than 0.5 were estimated to be female and individuals with probabilities greater than 0.5 were estimated to be male. A value of 0 would indicate an ambiguous or indeterminate sex.

There was one population-inclusive regression equation which was cross-validated by generating a hold-out-sample (HOS), where 30% of the study sample was removed from the BLR and a HOS population-inclusive BLR equation produced. This regression

equation was then tested on the 30% HOS and the accuracies and significance values reported.

In addition to the population-inclusive BLR equations, five population-specific BLR equations were produced for each population affinity group. The accuracies for these "original" models were reported. The "original" models were then tested on the study sample and the "applied" model accuracies and significance values were reported.

A BLR was run that combined both the metric and nonmetric data. A BLR was chosen in this instance because it avoided the problem of assuming linearity (Field 2009). This BLR produced a population-inclusive regression equation with statistically significant coefficient variables. This equation was cross-validated with a HOS regression where 30% of the population was held out and the remaining sample used to develop a regression which was then applied to the 30%. A BLR was run for each of the five population affinity groups, but results were not reported due to there being a lack of statistical significance for the coefficient variables.

Discriminant Function Analysis for Metric Model

The metric data were statistically analyzed using a discriminant function analysis in SPSS following Spradley and Jantz (2011) and because of the linear relationship cranial dimensions and measurements have (Field 2009: 599). Mahalanobis distance stepwise method was used to control for outlying variables. The sectioning points were obtained by taking the male and female mean and dividing by two. Values that fell above the sectioning point were considered male, values that fell below the sectioning point were

considered female, and values equal to the sectioning point were considered indeterminate or ambiguous sex.

There was one population-inclusive DFA equation developed with SPSS. The accuracies and significance values of this "original" model was reported. This "original" model was then tested on the study sample and new "applied" model accuracies produced. In addition to the population-inclusive DFA equation, five population-specific DFA equations were produced for each population affinity group. The accuracies for these "original" models were reported. The "original" models were then tested on the study sample in the same manner as with the population-inclusive DFA model.

Intra-Rater Reliability

Intra-rater reliability was calculated for 52 individuals, approximately 12.1% of the total study sample, who were scored and measured on a second occasion by the author. A random sampling procedure with SPSS was used to select the individuals that would be re-scored and remeasured.

The intraclass correlation coefficient (ICC) was calculated for the metric measurements in order to assess their reliability and relationship between variables. ICC measures the relationship between variables in the same class that measure the same thing (Field 2009:678), in this case variables that estimate sex. All metric measurements except one had ICC values above 0.8 and were statistically significant, indicating high intra-rater reliability

Cohen's kappa analysis (Cohen 1960, 1968) was employed to assess the degree of agreement or disagreement in both observations for the nonmetric traits of the cranium

and mandible. The kappa value evaluates the "proportion of agreement between observers corrected for chance and the standard measure of [intraobserver] reliability with nominal data (e.g., male vs. female)" (Walrath et al. 2004:135). Kappa values are one a scale between 0 to 1, with 0 representing a level of agreement that would be expected if ordinal scores were assigned at random (i.e., low agreement) and 1.0 representing perfect agreement (Walker 2005). The significance of the Kappa values was determined following Landis and Koch (1977) (Table 3.3). All Kappa values were between 0.365 and 0.563 and all traits were statistically significant.

Table 3.3. Agreement Measures for Kappa Statistics Adapted from Landis And Koch (1977).

Kappa Statistic	*Strength of Agreement*
<0.00	Poor
0.00-0.20	Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.81-1.00	Almost Perfect

CHAPTER FOUR: RESULTS

Current sex estimation models used by practitioners in the field of forensic anthropology are inherently reliant on specific populations (e.g., Spradley & Jantz 2011; Walker 2008) of U.S. Black, U.S. white, and Native American individuals that predominately come from university skeletal collections, like the Hamann-Todd and Terry skeletal collections (Klales 2020; Spradley et al. 2008; Tallman 2019; Tise et al. 2013). This phenomenon means that the conventional sex estimation methods are inherently exclusive and perform with lower classification accuracy when applied to populations that are dissimilar to the sample populations used in the development of the model. In addition to current sex estimation models being exclusive in nature, the ancestry estimation methods which they so greatly rely on have not figured proximate mechanisms or foundational theories for biological anthropology and human variation (e.g., evolutionary theory or ecogeographic) into their methodological development (Bethard & DiGangi 2020). This persistent oversight has led some forensic anthropologists to question the predictive value of the ancestry methods (Albanese & Saunders 2006; Bethard & DiGangi 2020; Blakey 2021; Carson 2006a; DiGangi & Bethard 2021), which in turn create a ripple effect for the sex estimation methods which rely on an identification of ancestry. The problematic foundation upon which conventional sex estimation models rest necessitates a population-inclusive model for estimating sex from the skeleton.

This chapter presents the results of the statistical analyses using SPSS. A binary logistic regression (BLR) analysis was run with the nonmetric data, a discriminant

function analysis (DFA) was run with the metric data, and another BLR was run to produce a model that combined the nonmetric and metric data. Intraobserver error was also calculated using Cohen's Kappa (1968) and Intraclass correlation coefficients for both the nonmetric and metric data collection. A number of equations were developed and cross-validated, both population-inclusive and population-specific, and the classification accuracies where generated and compared. Additionally, the population-specific models generated in isolation from the rest of the sample were applied to unselected populations to evaluate the statistical differences in accuracies.

They present research hypothesis proposed that (1) a population-inclusive model for estimating sex would produce classification accuracies that were not statistically different than the population-specific models and that (2) a population-inclusive model would perform better than the population-specific models when applied across populations.

The first hypothesis was supported by the results of the present study, which showed that the population-inclusive model did not perform statistically poorer than any of the population-specific models. The second hypothesis was partially supported as the results indicated that the population-inclusive model performed statistically better than some of the population-specific models. The results of this research indicate that a population-inclusive model can be applied in place of population-specific methods without hindering the estimation of sex as it will result in both statistically similar and statistically better classification accuracy.

Binary Logistic Regression of Nonmetric Traits

A BLR analysis was run with SPSS producing "original" population-inclusive and population-specific classification accuracies, including a hold-out-sample (HOS) to cross-validate the population-inclusive model "original" accuracies. The coefficients and constants for these models are presented in Table 4.1. The cross validated "original" accuracies for the BLR model are presented in Table 4.2. The model was then applied to the study sample, producing "applied" accuracies and these accuracies are presented in Table 4.3. The cross validated "applied" accuracies of population-inclusive model were compared to the cross-validated "applied" accuracies for the population-specific models to evaluate statistical significance. The statistical significance is noted in both accuracy tables (Table 4.2 and Table 4.3).

"Original" Accuracies for Population-Inclusive and -Specific Models

The cross-validated "original" accuracies are presented in Table 4.2. The population-inclusive model combined 3 morphological traits including the glabella, mastoid, and mental eminence. The population-inclusive "original" classification rate was 86.9% (females: 86.6%; males:87.1%). The population-inclusive HOS model combined 3 morphological traits including the glabella, mastoid, and nuchal crest. The population-inclusive HOS model had a total "original" classification rate of 84.9% (females: 88%; males:82.4%)**.**

The population-specific African American model combined 2 morphological traits including the glabella and the mastoid. The African American model had a total "original" classification rate of 82.2% (females: 86.8%; males:77.1%). The population-

specific Asian American model combined 2 morphological traits including the glabella and the mastoid. The Asian American model had a total "original" classification rate of 91% (females: 84.2%; males:93.2%). The population-specific European American model combined 2 morphological traits including the glabella and the mental eminence. The European American model had a total "original" classification rate of 85.2% (females: 87.5%; males:82.9%). The population-specific Latin American model combined 2 morphological traits including the glabella and the nuchal crest. The Latin American model had a total "original" classification rate of 84.6% (females: 89.2%; males:80.5%). The population-specific Native American model combined 2 morphological traits including the glabella and the mastoid process. The Native American model had a total "original" classification rate of 88.6% (females: 92.1%; males:85.4%).

"Applied" Accuracies for Population- Inclusive and Exclusive Models

The population-inclusive model produced an overall "applied" accuracy of 81% (females:90.7%; males:73.3%). The population-inclusive HOS model produced an overall "applied" accuracy of 83.2% (females:91.4%; males:76.7%). The African American model produced an overall "applied" accuracy of 79.7% (females:88.1%; males:70.3%). The overall "applied" accuracy for the Asian American model was 91.7% (females:84.2%; males:93.8%). The European American model produced an overall "applied" accuracy of 85.2% (females:87.5%; males:82.9%). The Latin American model produced an overall "applied" accuracy of 78.8% (females:90.5%; males:67.4%). The Native American model produced an overall "applied" accuracy of 88.2% (females:92.7%; males:84.1%).

Pearson's Chi-Square Test of Significance

Using Pearson's chi-square statistic, with a *p*-value of 0.05, the statistical significance of the both the "original" and "applied" classification accuracies was tested against the population-inclusive "original" and "applied" classification accuracies. All "original" classification accuracies had *p*-values greater than 0.5 and were not significantly different than the population-inclusive model for the "original" classification accuracies. The population-inclusive model had an overall "applied" accuracy of 81%. The African American, European American, Latin American, and Native American "applied" accuracies had *p*-values greater than 0.5 and were not significantly different from the population-inclusive "applied" accuracy. The Asian American "applied" overall accuracy (91.7%) had a *p*-value less than 0.5 and was significantly different than the population-inclusive "applied" accuracy.

Table 4.1. Binary Logistic Regression Analysis Models for Inclusive Population and Specific Populations.

Equation	***Classification Function with Stepwise Selected Variable****
Population-Inclusive	(glabella score*1.385)+(mastoid score*0.902)+(mental eminence score*0.44)+(-5.888)
Population-Inclusive**	(glabella score*1.363)+(mastoid score*0.876)+(nuchal crest score*0.393)+(-5.664)
African American	(glabella score*1.335)+(mastoid score*1.046)+(-5.164)
Asian American	(glabella score*3.033)+(mastoid score*1.012)+(-6.438)
European American	(glabella score*1.628)+(metal eminence score*1.002)+(-6.309)

Latin American	(glabella score*1.324)+(nuchal crest score*0.995)+(-5.18)
Native American	(glabella score*1.827)+(mastoid score*1.276)+(-7.037)

***All Class Means Significantly Different, P< 0.05.**

**** Population-Inclusive 70% Hold-Out Sample.**

Table 4.2. Cross-Validated "Original" Classification Rates for Binary Logistic Regression Models.*

Equation	*Female n*	*Male n*	*Female %*	*Male %*	*Overall %***
Population-Inclusive Model	149	189	86.6%	87.1%	87%
Population-Inclusive Model***	103	122	88.0%	82.4%	85%
African American Model	33	27	86.8%	77.1%	82%
Asian American Model	16	61	84.2%	93.2%	91%
European American Model	35	34	87.5%	82.9%	85%
Latin American Model	33	33	89.2%	80.5%	85%
Native American Model	35	35	92.1%	85.4%	89%

***Cut Value Is 0.5**

****All Classification Accuracies are Not Significantly Different than the Population-Inclusive Model (P Value <0.5; (2-Tailed)**

*****Population-Inclusive 70% Hold-Out Sample**

Table 4.3. "Applied" Classification Accuracies for Binary Logistic Regression Models.

Equation	*Female n*	*Male n*	*Female %*	*Male %*	*Overall %*
Population-Inclusive Model	156	159	90.7%	73.3%	81.0%
Population-Inclusive Model**	53	56	91.4%	76.7%	83.2%
African American Model	37	26	88.1%	70.3%	79.7%
Asian American Model	16	61	84.2%	93.8%	91.7%*
European American Model	35	34	87.5%	82.9%	85.2%
Latin American Model	38	29	90.5%	67.4%	78.8%
Native American Model	38	37	92.7%	84.1%	88.2%

***Significantly Different Accuracy Compared to the Overall % for the Population-Inclusive Model,** *P*-Value **<0.05 (2-Tailed)**

****Population-Inclusive 70% Hold-Out Sample**

Discriminant Function Analysis of Metric Measurements

A DFA was run with SPSS producing one population-inclusive and five separate population-specific models for estimating sex. These DFA equations are presented in Table 4.4. The "original" population-inclusive and population-specific cross-validated classification accuracies for this model are presented in Table 4.5. The DFA equations were then applied to the study sample, producing "applied" classification accuracies, presented in Table 4.6. Using Pearson's chi-square statistics, the population-specific classification rates from both the "original" and "applied" tests were compared "original" and "applied" accuracies for the population-inclusive model to evaluate statistical significance.

"Original" Accuracies for Population- Inclusive and Exclusive Models

The population-inclusive model combined ten points of measurement, including glabella occipital length, bizygomatic breadth, biauricular breadth, minimum frontal breadth, nasal height, orbital height, mastoid height, bigonial breadth, maximum ramus height, and mandibular length. The population-inclusive model had a total cross-validated "original" classification rate of 86.7% (females: 88%; males:85.7%). The population-specific African American model combined 6 points of measurement, including bizygomatic breadth, biauricular breadth, minimum frontal breadth, mastoid height, bicondylar breadth, and maximum ramus height. The African American model had a total cross-validated "original" classification rate of 86.1% (females: 84.2%; males:88.2%). The population-specific Asian American model combined 3 points of measurement, including basion-nasion length, frontal chord, and mastoid height. The Asian American

model had a total cross-validated "original" classification rate of 88.2% (females: 93.8%; males:86.7%). The population-specific European American model combined 5 points of measurement, including bizygomatic breadth, orbital height, mastoid height, maximum ramus height, and mandibular length. The European American model had a total cross-validated "original" classification rate of 81.0% (females: 81.6%; males:80.5%). The population-specific Latin American model combined 2 points of measurement, including bizygomatic breadth and maximum ramus height. The Latin American model had a total cross-validated "original" classification rate of 77.1% (females: 82.9%; males:71.4%). The population-specific Native American model combined 5 points of measurement, including glabella occipital length, orbital height, mastoid height, bigonial breadth, and maximum ramus height. The Native American had a total cross-validated "original" classification rate of 86.3% (females: 80.6%; males:90.9%).

"Applied" Accuracies for Population- Inclusive and Exclusive Models

The population-inclusive model produced an overall "applied" accuracy of 87% (females:90.4%; males:84.3%). The African American model produced an overall "applied" accuracy of 95% (females:89.5%; males:100%). The Asian American model produced an overall "applied" accuracy of 91% (females:93.8%; males:90%). The European American model produced an overall "applied" accuracy of 86% (females:84.2%; males:87.8%). The Latin American model produced an overall "applied" accuracy of 78% (females:85.4%; males:71.4%). The Native American model produced an overall "applied" accuracy of 73% (females:80.6%; males:67.4%).

Pearson's Chi-Square Test of Significance

Using Pearson's chi-square statistic, with a *p*-value of 0.05, the statistical significance of the both the "original" and "applied" classification accuracies was tested against the population-inclusive "original" and "applied" classification accuracies. The population-inclusive "original" classification accuracy was 77.1%. The African American, Asian American, European American, and Native American "original" accuracies all had *p*-values greater than 0.5 and were not significantly different than the population-inclusive "applied" classification accuracy. The Latin American "original" classification accuracy was 77.1% and had a statistically significant *p*-value of less than 0.5. The population-inclusive model had an overall "applied" accuracy of 87%. All of the population-specific "applied" accuracies were not significantly different from the population-inclusive "applied" accuracies except for the Latin American "applied model which had a *p*-values less than 0.05. The Latin American population-specific model produced an "applied" accuracy of 78%.

Table 4.4. Discriminant Function Analysis Models for Inclusive Population and Specific Populations.

Equation	***Classification Function with Stepwise Selected Variable***
Population-Inclusive	(glabella occipital length*0.057)+(bizygomatic breadth*0.126)+(biauricular breadth*-0.047)+(minimum frontal breadth*-0.069)+(nasal height*0.059)+(orbital height*-0.115)+(mastoid height*0.081)+(bigonial breadth*0.037)+(maximum ramus height*0.074)+(mandibular length*-0.046)+(-20.182)
African American	(bizygomatic breadth*0.335)+(biauricular breadth*-0.188)+(minimum frontal breadth*-0.185)+(mastoid height*0.123)+(bicondylar breadth*-0.089)+(maximum ramus height*0.185)+(-9.561)

Asian American	(basion-nasion length*0.142)+(frontal chord*0.102)+(mastoid height*0.101)+(-29.68)
European American	(bizygomatic breadth*0.14)+(orbital height*-0.337)+(bigonial breadth*0.079)+(maximum ramus height*0.109)+(mandibular length*-0.085)+(-13.013)
Latin American	(bizygomatic breadth*0.136)+(maximum ramus height*0.106)+(-24.507)
Native American	(glabella occipital length*0.082)+(orbital height*-0.197)+(mastoid height*0.082)+(bigonial breadth*0.102)+(maximum ramus height*0.075)+(-25.132)

***All Class Means Significantly Different, P<0.05**

Table 4.5. Cross-Validated "Original" Classification Rates for Discriminant Function Models.

Equation	*Female n*	*Male n*	*Female %*	*Male%*	*Overall %*
Population-Inclusive Model	146	180	88.0%	85.7%	86.7%
African American Model	32	30	84.2%	88.2%	86.1%
Asian American Model	15	52	93.8%	86.7%	88.2%
European American Model	31	33	81.6%	80.5%	81.0%
Latin American Model	34	30	82.9%	71.4%	77.1%*
Native American Model	29	40	80.6%	90.9%	86.3%

***Significantly Different Accuracy Compared to the Population-Inclusive Model, *P*-Value <0.05 (2-Tailed).**

Table 4.6. "Applied" Classification Accuracies for Discriminant Function Models.

Equation	*Female n*	*Male n*	*Female %*	*Male %*	*Overall %*
Population-Inclusive Model	150	177	90.4%	84.3%	87%
African American Model	34	35	89.5%	100.0%	95%
Asian American Model	15	54	93.8%	90.0%	91%
European American Model	32	36	84.2%	87.8%	86%
Latin American Model	35	30	85.4%	71.4%	78%*
Native American Model	29	42	80.6%	95.5%	89%

***Significantly Different Accuracy Compared to the Population-Inclusive Model, *P*-Value <0.05 (2-Tailed)**

Combined Binary Logistic Regression of Metric And Nonmetric Traits

A BLR analysis was run with SPSS that combined the metric and nonmetric data (i.e., combined BLR) to produce a population-inclusive sex estimation model (Table 4.7). A HOS BLR was also run to cross-validate the population-inclusive model. The coefficients and constants for these models are presented in Table 4.7. The cross validated "original" accuracy for the BLR model was recorded and all class means were statistically significant ($p<0.05$) (Table 4.8). The population-inclusive model was then applied to the study sample, producing an "applied" accuracy (Table 4.9). The significance values for the population-specific class coefficients for each specific population affinity all consistently exceeded the established significance value of 0.05, and were not statistically significant.

Table 4.7. Population-Inclusive Combined (Nonmetric and Metric) Binary Logistic Regression Analysis Models.

Equation	*Classification Function with Stepwise Selected Variable**
Population-Inclusive	(glabella score*1.13)+(mastoid score*0.957)+(mental eminence score*0.594)+(glabella occipital length*0.102)+(bizygomatic breadth*0.1620+(maximum ramus height*0.147)+(mandibular length*-0.101)+(-44.921)
Population-Inclusive**	(glabella score*2.027)+(bizygomatic breadth*0.263)+(-38.097)

***All Class Means Significantly Different, *P*<0.05.**

****Population- Inclusive 70% Hold-Out Sample.**

"Original" and "Applied" Accuracies

The population-inclusive model combined 7 morphological and metric traits including the glabella, mastoid, mental eminence, glabella occipital length, bizygomatic breadth, maximum ramus height, and mandibular length (Table 4.7). The population-inclusive model had a total "original" classification rate of 91.6% (females: 88.8%; males:93.3%) and an overall "applied" accuracy of 88.8% (females:88.1%; males:89.3%).

The population-inclusive HOS model combined 2 traits including the glabella and bizygomatic breadth (Table 4.7). The population-inclusive HOS model had a total "original" classification rate of 87.6% (females: 84.2%; males:89.8%) and an overall "applied" accuracy of 88.8% (females:91.2%; males:86.6%).

Table 4.8. "Original" Cross-Validated Classification Rates for Combined (Nonmetric and Metric) Binary Logistic Regression Models.

Equation	*Female n*	*Male n*	*Female %*	*Male %*	*Overall %*
Population-Inclusive Model	111	182	88.8%	93.3%	91.6%
Population-Inclusive Model*	32	53	84.2%	89.8%	87.6%

***Population-Inclusive 70% Hold-Out Sample.**

Table 4.9. Population-Inclusive "Applied" Classification Accuracies for Combined (Nonmetric and Metric) Binary Logistic Regressions*.

Equation	*Female n*	*Male n*	*Female %*	*Male %*	*Overall %*
Population-Inclusive Model	141	160	88.1%	89.3%	88.8%
Population-Inclusive Model**	52	57	91.2%	86.8%	88.8%

***Cut Value 0.5.**

****Population-Inclusive 70% Hold-Out Sample.**

Intra-Rater Reliability

Intraobserver agreement was evaluated for both the nonmetric and metric data collection. On a separate occasion, a random sample of approximately 12% (n=52) of the study sample was revisited for nonmetric and metric data collection to assess intraobserver agreement.

Nonmetric Trait Scoring

Intraobserver error for nonmetric scoring was assessed using Cohen's Kappa statistic (Cohen 1968) in order to evaluate the magnitude of potential error and level of consistency and level of agreement was established following Landis and Koch (1977). The nuchal crest performed with a fair level of agreement while the mastoid, supraorbital margin, glabella, and mental eminence all performed with a moderate level of agreement (Table 4.10).

Metric Measurements

The intra-class correlation coefficients (ICC) ranged from 0.777 to 0.989 with the glabella occipital length performing the best and the parietal chord performing the worst. All class means were statistically significant, with *p*-values below 0.001.

Table 4.10. Intraobserver Error Rates for Nonmetric Trait Scores.

*Morphological Traits**	*Kappa Value*	*Level of Agreement***	*Asymptotic SE*	*Approximate T*	*Approximate significance*
Nuchal Crest	0.365	Fair	0.081	5.839	<0.001
Mastoid	0.563	Moderate	0.088	7.267	<0.001
Supraorbital Margin	0.432	Moderate	0.088	6.097	<0.001
Glabella	0.531	Moderate	0.083	7.393	<0.001

Mental Eminence	0.452	Moderate	0.088	6.21	<0.001

***All Class Means Significantly Different, P<0.001**

****Level of Agreement Following Landis And Koch (1977)**

Table 4.11. Intraobserver Error Rates for Metric Measurements.*

Metric Measurement	*Valid cases n*	*Valid cases %*	*Excluded cases n*	*Total cases n*	*ICC (for average measures)*[a]	*95% Confidence Interval*
Glabella Occipital Length	51	98.10%	1	52	0.989	0.98-0.994
Bizygomatic Breadth	52	100%	0	52	0.931	0.88-0.96
Basion Bregma Height	46	88.50%	6	52	0.825	0.682-0.903
Basion Nasion Length	51	98.10%	1	52	0.923	0.865-0.956
Biauricular Breadth	50	96.20%	2	52	0.909	0.836-0.949
Nasion Prosthion Height	38	73.10%	14	52	**0.981**	0.961-0.99
Minimum Frontal Breadth	52	100%	0	52	0.88	0.792-0.931
Nasal Height	51	98.10%	1	52	0.941	0.896-0.966
Orbital Height	52	100%	0	52	0.93	0.877-0.96
Frontal Chord	47	90.40%	5	52	0.908	0.836-0.949
Parietal Chord	38	73.10%	14	52	**0.777**	0.575-0.883
Occipital Chord	39	75%	13	52	**0.926**	0.859-0.961
Foramen Magnum Breadth	51	98.10%	1	52	0.98	0.965-0.988
Mastoid Height	51	98.10%	1	52	0.837	0.509-0.929
Bigonial Breadth	52	100%	0	52	0.987	0.978-0.003
Bicondylar Breadth	52	100%	0	52	0.938	0.893-0.965
Maximum Ramus Height	52	100%	0	52	0.945	0.904-0.968
Mandibular Length	50	96.20%	2	52	0.904	0.813-0.948

***All Class Means Significantly Different, P<0.001.**

[A] Those in Bold Excluded from The Discriminant Function Analysis.

Population-Specific Model Performance and Accuracy

The population-specific models were tested on the populations not included in the original development of the population-specific model (e.g., African American model tested on Latin American population). These tests produced accuracy rates which were then compared to the population-inclusive "applied" accuracies to evaluate whether or

not there was a statistical difference in accuracy. The descriptive statistics for these BLR accuracies are presented in Table 4.12 and the statistics for the DF analysis are presented in Table 4.13.

Binary Logistic Regression

The BLR population-specific European American, Latin American and Native American models all produced "applied" accuracy rates that were significantly different from the population-inclusive "applied" accuracy when applied to the Asian American population sample (Table 4.12).

Discriminant Function Analysis

The African American, Asian American, European American, Latin American, and Native American DF population-specific models all produced at least one "applied" accuracy rate that was significantly different from the population-inclusive "applied" accuracy (Table 4.13).

Table 4.12. "Applied" Classification Accuracies for Binary Logistic Regression Population-Specific Models as Tested on Unselected Populations.

Applied Model	*Test Population*	*Female n*	*Female %*	*Male n*	*Male %*	*Overall %*
African American	Asian Am.	19	100%	52	80%	84.5%
	European Am.	34	82.9%	37	86%	84.5%
	Latin Am.	39	92.9%	32	74.4%	83.5%
	Native Am.	38	92.7%	38	86.4%	89.4%
Asian American	African Am	30	71.4%	32	86.5%	78.5%
	European Am.	19	46.3%	43	100%	73.8%
	Latin Am.	25	59.5%	37	86%	72.9%
	Native Am.	27	65.9%	43	97.7%	82.4%

European American	African Am	36	94.7%	21	60%	78.1%
	Asian Am.	19	100%	34	57.6%	67.9%*
	Latin Am.	33	89.2%	27	65.9%	76.9%
	Native Am.	33	86.8%	35	85.4%	86.1%
Latin American	African Am	38	90.5%	24	64.9%	78.5%
	Asian Am.	19	100%	36	55.4%	65.5%*
	European Am.	36	87.8%	35	81.4%	84.5%
	Native Am.	36	87.8%	38	86.4%	87.1%
Native American	African Am	39	92.9%	21	56.8%	75.9%
	Asian Am.	19	100%	37	56.9%	66.7%*
	European Am.	36	87.8%	33	76.7%	82.1%
	Latin Am.	39	92.9%	27	62.8%	77.6%

***Significantly Different "Applied" Accuracy than Population-Inclusive "Applied" Accuracy, *P*-Value <0.05**

Table 4.13. "Applied" Classification Accuracies for Discriminant Function Analysis Population-Specific Models as Tested on Unselected Populations.

Applied Model	*Test Population*	*Female n*	*Female %*	*Male n*	*Male %*	*Overall%*
African American	Asian Am.	15	88.2%	38	67.9%	72.6%*
	European Am.	35	92.1%	24	58.5%	74.7%*
	Latin Am.	38	92.7%	23	54.8%	73.5%*
	Native Am.	29	80.6%	29	70.7%	75.3%*
Asian American	African Am	14	63.6%	29	85.3%	76.8%*
	European Am.	25	69.4%	36	85.7%	78.2%*
	Latin Am.	26	78.8%	29	74.4%	76.4%*
	Native Am.	22	78.6%	35	81.4%	80.3%
European American	African Am	39	97.5%	17	50%	75.7%*
	Asian Am.	10	58.8%	45	81.8%	76.4%*
	Latin Am.	33	82.5%	28	66.7%	74.4%*
	Native Am.	26	68.4%	38	90.5%	80%
Latin American	African Am	32	80%	29	78.4%	79.2%*
	Asian Am.	11	64.7%	51	86.4%	81.6%
	European Am.	35	92.1%	32	76.2%	83.8%
	Native Am.	21	55.3%	40	93%	75.3%*
Native American	African Am	27	69.2%	31	88.6%	78.4%
	Asian Am.	15	88.2%	51	91.1%	90.4%

	European Am.	26	65%	38	90.5%	78%*
	Latin Am.	29	70.7%	35	81.4%	76.2%*

***Significantly Different "Applied" Accuracy than Population-Inclusive "Applied" Accuracy, *P*-Value <0.05.**

CHAPTER FIVE: DISCUSSION

This chapter discusses the results presented in the previous chapter including the performance accuracy for the population-inclusive and population-specific models. Additionally, the statistical analyses and accuracies for the BLR nonmetric model, DFA metric model, the BLR nonmetric and metric model, and the degree of intra-rater reliability and the effect that intraobserver error had on the overall analysis and classification accuracies will be discussed. This chapter will also consider the application of the population-specific BLR and DFA models across the five different population groups. The population-specific models, when applied across sample populations, performed more poorly when compared with population-inclusive models.

The results of this study indicate that the initial hypothesis is valid and that a population-inclusive method can be employed to accurately estimate sex without producing significantly different or, more importantly, statistically lower classification rates. This is particularly important with current discussions in the field of forensic anthropology regarding the problematic nature of ancestry estimation in the biological profile (Albanese & Saunders 2006; Blakey 2021; Bethard & DiGangi 2020; DiGangi & Bethard 2021; Tallman et al. 2021), and suspending ancestry estimation methods until there is a better understanding of the primary mechanisms of population variation (Bethard & DiGangi 2020). Additionally, while it is clear that population variation and differences do exist, the results of this study indicate that the effect on sex estimation is minimal. In cases where ancestry is unknown or is intentionally not estimated, a population-inclusive model for estimating sex would be the most appropriate method.

Nonmetric Traits

The nonmetric traits used in the population-inclusive BLR model included the glabella, mastoid process, and mental eminence. The population-inclusive model incorporated more nonmetric traits than the population-specific models (three instead of two), based on the statistical significance of each individual variable. This is likely due to the variation in sample sizes as the population-inclusive model was developed on 389 individuals, whereas the population-specific models were all developed on population size between 73 and 81 individuals. Sample size is often one of the greatest limitations in stepwise procedures (Field 2009).

Each nonmetric BLR model (both population-inclusive population-specific) included the glabella score in the regression, and the African American, Asian American, and Native American models also all included the mastoid. Research suggests that the glabella and mastoid are the best discriminators of sex due to their highly sexually dimorphic nature (Tallman 2019; Walker 2008). Tallman (2019) found that the glabella and mastoid performed are the best at estimating sex on a population of an East and Southeast Asian-derived population. Garvin et al. (2014) similarly found that when estimating sex in Arikara Native Americans, Nubian, U.S. Black and U.S. white populations the mastoid process and glabella performed the best.

Both the population-inclusive and the European American model incorporated the mental eminence into the regression analysis. Despite previous research suggesting that a difficulty exists with scoring the mental eminence (Garvin et al. 2014; Lewis & Garvin 2016; Williams & Rogers 2006), the mental eminence performed with moderate intra-

observer agreement and with statistically significant class means for the population-inclusive BLR model as well as the population-specific European American model. The Latin American model was the only population-specific model that incorporated the nuchal crest into the regression equation, despite the nuchal crest performing with only a fair level of intra-observer agreement.

Classification Accuracy of Nonmetric Traits

Overall, the population-inclusive model produced classification accuracies that ranged from 81.0-87%, and the population-specific models produced classification accuracies that ranged from 78.8-91.7%. The population-inclusive BLR model did not have significantly lower classification accuracy rates than the population-specific models. This indicates that a population-inclusive model for estimating sex with nonmetric traits of the skull could be used with sufficient success in place of population-specific methods for estimating sex.

The Asian American model produced the highest accuracy rate that ranged from 91.0-91.7% compared to the other population-specific models. The "original" accuracy rate saw a 4% increase relative to the population-inclusive model, but this was not statistically significantly higher. On the other hand, the "applied" accuracy rate (91.7%) saw a significant 10.7% increase in accuracy. This increase in accuracy is potentially skewed due to the disproportionately small female sample sizes. Research indicates that sexual dimorphism in Asian-derived populations is reduced (Roth et al. 2013; Tallman 2019), so the Asian American population-specific models should demonstrate lower accuracies and higher sex biases (Tallman 2019). Instead, the Asian American BLR model produced a

higher accuracy rate for males which indicates the statistical significance is sample-size related (Field 2009). Additionally, it is likely that the individuals in the Asian American sample lived in the U.S., with access to relatively high caloric food and nutrition and research shows that groups with high levels of nutrition (i.e., Americans) tend to exhibit more sexual dimorphism overall, regardless of continental ancestry (Tallman 2019). Populations that are nutrient deficient exhibit slower maturation rates and reduced sexual dimorphism (Stinson 1985, 2012).

The Latin American model had the lowest overall "applied" accuracy rate for the BLR model (78.8%), but this was not statistically significantly lower than the population-inclusive model, but had the lowest "applied" accuracy compared to the other population-specific models and the population-inclusive model. There was only a 2% reduction in the "original" accuracy relative to the population-inclusive model. Similarly, there was a 2.2% reduction in "applied" accuracy for the population-specific Latin American model, which can be attributed to the low "applied" accuracy for males (67.4%). This was also the lowest "applied" accuracy rate for females. The "applied" accuracy rate for males using the Latin American model was the lowest percentage (67.4%) relative to the other "applied" male accuracies and compared to the "applied" Latin American female accuracy at 90.5%. This is potentially due to a higher variability in skull morphology for Latin American-derived skulls, and especially males. Klales and Cole (2017) found that Latin American male skulls were more variable than the females in score frequency for all traits except the mental eminence. Males received variable scores for the nuchal crest especially, with the majority being given a score of 2 (on a scale of 1 to 5) (Klales & Cole

2017). The majority of males were also given scores of 2 for the glabella (Klales & Cole 2017).

The African American model had an overall classification accuracy that ranged from 79.7-82.0%. The "original" classification accuracy (82.0%) saw a 5.0% reduction relative to the population-inclusive model and the "applied" accuracy (79.7%) saw a 1.3% reduction. While neither accuracy was statistically different from the population-inclusive model, the African American model did have the second lowest "applied" classification accuracy.

The European American model overall classification accuracy ranged from 85.0-85.2% The "original" classification accuracy (85.0%) saw a 2.0% reduction in accuracy relative to the population-inclusive model and the "applied" accuracy (85.2%) saw a 4.2% increase relative to the population-inclusive model, but neither was significantly different.

The Native American model had an overall classification accuracy that ranged from 88.2-89.0%. The "original" classification accuracy (89.0%) saw a 2% increase relative to the population-inclusive model, and the "applied" accuracy (88.2%) saw a 7.2% increase in accuracy relative to the population-inclusive model, but neither was significantly different.

Walker's study (2008) found that the population-specific BLR equation for African American and European American individuals correctly classified individuals 84-88% of the time. Similarly, Garvin et al. (2014) found that population-specific BLR equations for

Arikara Native American, medieval Nubian, U.S. Black, and U.S. white individuals were correctly classified 74-99% of the time, which is consistent with the results of this study.

Potential Sex Bias

The "applied" accuracies for both the population-inclusive and population-specific nonmetric models all show higher "applied" classification rates for females than males, except for the Asian American model which was likely affected by sample size. Overall, this pattern with lower "applied" male accuracy indicates that the BLR models may be biased toward females, in that they are more accurate in classifying females. Other research has found a similar sex bias in that females are more often classified correctly than males especially in more modern samples, due either to population variation (Klales & Cole 2017; Tallman & Go 2018) or secular change.

Metric Traits

The population-inclusive model incorporated the most variables in the final DFA, which is attributed to the overall sample size (Field 2009). No DFA model used the same combination of variables. The Latin American model used the smallest number of variables (two) indicating that either there were not enough measurements to choose from (i.e., small sample size) or they were not statistically significant (Field 2009). Glabella occipital length, bizygomatic breadth, biauricular breadth, mandibular length, maximum ramus height, orbital height, minimum frontal breadth, bigonial breadth, and mastoid height are used in more than one equation. Richard et al. (2014) found that glabella occipital length, bizygomatic breadth, and biauricular breadth, in particular, all were

reliable landmarks for discriminating sex in a sample of skulls from the Bass skeletal collection.

Classification Accuracy of the Metric Traits

The population-inclusive DFA model does not have significantly lower classification accuracy rates than any of the population-specific models. This, likewise indicates that metric population-inclusive methods for sex estimation can be used with statistical accuracy in the case of unknown ancestry, no existing population-specific method, or where ancestry is not estimated. Furthermore, the "applied" accuracy rates for the metric traits for the population-inclusive and population-specific models were higher than their "original" counterparts which is likely due to an increase in sample size (Field 2009).

The DFA model with the highest accuracy rate was the population-specific Asian American model, but this was not significantly higher than the population-inclusive model. The Asian American model also had the highest female accuracy rate. Research suggests that this is likely due to the reduced sexual dimorphism found in Asian-derived populations, which is generally evidenced by high accuracy rates biased towards females (Tallman & Go 2018; Tallman 2019).

The Latin American model had the overall lowest "original" and "applied" accuracy compared to the other population-specific models and was significantly lower than the population-inclusive model. In particular, the Latin American males had the lowest "original" and "applied" accuracy. Research indicates that this could be due to a high level of skeletal variation in male skulls specifically (Klales & Cole 2017). A metric study on the estimation of sex from the FDB using Hispanic skeletons by Spradley et al.

(2008) found that females were nearly always classified correctly, whereas "too many males [were] classified as female" (Spradley et al. 2008:26).

Combined Nonmetric and Metric Traits

Classification Accuracy of Nonmetric and Metric Traits

Overall, the population-inclusive model that combined both nonmetric scores and metric measurements had higher "original" and "applied" accuracies (88.8-91.6%) than both the separate nonmetric and metric population-inclusive model "original" and "applied" accuracies.

The combined model did not have a significantly different accuracy than the metric model "original" and "applied" accuracies or the nonmetric "original" accuracy, but it was significantly more accurate than the nonmetric model "applied" accuracy (increase of 10.7%). A BLR was attempted through SPSS, but none of the population-specific models yielded class means with *p*-values greater than 0.05 and therefore were not statistically significant.

Most sex estimations conventionally used in forensic casework apply nonmetric and metric models in a complementary fashion (Klales 2020) but there does not appear to be any methods currently employed that combine both the nonmetric and metric traits into one regression equation as seen in this study. Since most practitioners prefer to use both nonmetric and metric methods estimating sex with the skull, this regression that combines both could be a valuable piece of the forensic anthropologists' toolkit. Additionally, since the classification accuracies for this combined equation are higher

than the nonmetric and metric models alone, it is recommended that more studies be done in order to explore the validity of this type of model.

Intra-Rater Reliability

On a separate occasion, a random sample of approximately 12% (n=52) of the study sample was revisited for nonmetric and metric data collection to assess intraobserver agreement.

Nonmetric Trait Scoring

Intraobserver agreement was highly significant for all nonmetric trait scores (*p*-values <0.001). Following Landis and Koch (1977), the nuchal crest performed with a higher level of variability (fair level of agreement) in scoring than the mastoid process, supraorbital margin, glabella, and mental eminence. All the traits except for the nuchal crest performed with a moderate level of agreement. The nuchal crest and supraorbital margin had the lowest kappa values (0.365 and 0.432, respectively), suggesting that both traits were difficult to score consistently between observations. These results are dissimilar to the intraobserver reliability test from Tallman's (2019) cranial nonmetric analysis on East and Southeast Asian individuals where the nuchal crest had a substantial level of agreement between observations. However, the mastoid process, supraorbital margin, glabella, and mental eminence all had a moderated level of agreement between observations (Tallman 2019). Garvin et al. (2014) found that the nuchal crest and metal eminence had the lowest level of agreement. Garvin et al. (2014) postulated that the nuchal crest was more difficult to score consistently because it involves a larger area relative to the other morphological traits and the range of shape variation is relatively

smaller and thus more difficult to visualize. Even though the 3D CT image was manipulated and rotated in order to examine the nuchal crest from all angles, the visual extent of rugosity could have been inhibited due to the quality of the CT image or PVE. Stull et al. (2014) found that the ability to assess volume-rendered CT images was impacted by PVE during the volume rendering process. This occurs when the CT scanner has difficulty in distinguishing between materials with different Hounsfield units, such as air and bone, and results in the appearance of artifacts that can obscure an area or softening of the object's surface (Stull et al. 2014). In the case of the nuchal crest, the PVE phenomenon could have reduced the appearance of the rugosity making it more difficult to score consistently. In the case of the supraorbital margin, the difficulty lay in the inability to estimate the relative width and roundness or sharpness of the margin on a computer image. In the future, caution should be applied when analyzing the nuchal crest and supraorbital margin via volume-rendered images.

The mastoid process and glabella performed with moderate levels of agreement, but had the highest kappa values (0.563 and 0.531 respectively) comparatively. Walker (2008) similarly found that the glabella and mastoid had the highest levels of agreement relative to the nuchal crest, supraorbital margin, and mental eminence.

Metric Measurements

The ICC values range from 0.777 to 0.989 which indicates a high level of intra-rater reliability for the metric measurements. This affirms that measurements can be successfully taken on CT scans and perform better than nonmetric traits. The glabella occipital length has the highest ICC (0.989) and the parietal chord had the lowest ICC

(0.777). The parietal chord, occipital chord, and nasion-prosthion height all had the most cases excluded from the ICC analysis so they were also removed from the DFA in order to maximize sample size. These traits were likely excluded from the ICC due to a lack of measurements. One of the artifacts of PVE is a softening of overall images that can cause cranial sutures, especially, to appear faded or obliterated (Stull et al. 2014), making cranial chords difficult to measure. The nasion prosthion height was often not able to be measured to due alveolar resorption, damage to craniometrics landmark, or PVE, potentially from dental fillings or soft tissue. A study by Menéndez (2017) looked at intraobserver measurement error of 3D craniofacial landmarks and found that nasion was one of the craniofacial landmarks that had the highest intraobserver error rates and was difficult to establish consistently across multiple observational sessions. The occipital chord (Herrera and Tallman 2019; Richard et al. 2014) and parietal chord (Richard et al. 2014) are also craniometrics landmarks that have been shown to have variable intraobserver reliability. The cranial chords rely on landmarks demarcated by suture lines which could also be affected by the quality of the 3D volume-rendered image (Richard et al. 2014). In the future, it would be best to remove measurements that include cranial sutures because of the difficulty existing in scoring them consistently when they are faded or obliterated due to PVE or CT quality. Removing these measurements may help to maximize sample size.

Population-Specific Model Performance

The population-specific models, when applied across sample populations, produced overall lower accuracy rates for both the nonmetric and metric traits compared to the

population-inclusive model and the population-specific models when applied to the appropriate sample population. The results from applying the population-specific model across the five sample population affinities further emphasizes the need for researchers to use caution when applying sex estimation models that have been developed from specific populations to different population groups.

Nonmetric Equation

On average, the Native American sample population had the highest classification rates when the other population-specific models (e.g., African American, European American, Asian American, and Latin American) were applied to it. The Asian American group was the only sample population that consistently had significantly low overall classification rates. The female accuracy was consistently 100%, but the male accuracy ranged from 55.4-80.0%. This is likely due to the small sample size, as well as the reduced sexual dimorphism and inherent sex bias that accompanies it.

Metric Measurements

The metric population-specific models produced lower overall accuracy rates compared to the nonmetric population-specific models. The African American model, when applied to the other four sample groups, produced significantly reduced accuracy rates compared to the population-inclusive model. The European American model, likewise, produced significantly reduced accuracy rates when applied to every sample group except Native American. The Latin American model produced significantly reduced classification accuracy rates when applied to the African American group and the

Native American group. Finally, the Native American model produced significantly reduced classification accuracy rates when applied to the European American and Latin American groups.

Data Collection from Volume-Rendered CT Images

The majority of nonmetric traits had a moderate level of intraobserver agreement, while the metric traits, as aforementioned, all had statistically significant ICC values and overall higher accuracy rates. The nonmetric data collection was more difficult compared to the metric data collection. The data collection for the nonmetric scores, that being visual examination of volume-rendered CT images of the skull from multiple angles, was impacted by the nature of viewing an image on a computer rather than holding it in your hand as well as PVE. While research shows that data collection from volume-rendered CT images of the skeleton can be accomplished with statistically significant accuracy (Decker et al. 2011; Robinson et al. 2008; Stull et al. 2014), certain traits are more difficult to examine than others.

The supraorbital margin was the most difficult, despite performing with a moderate level of intraobserver agreement. The most difficult part of scoring the supraorbital margin was distinguishing between the intermediate ordinal scores (i.e., 2-3 and 3-4). A study that examines the supraorbital margin more critically would be interesting. This study could either be a seriation-type approach, where a sample of volume-rendered CT scans of the skull are used and supraorbital margin isolated and placed in order based on gracility (and sharpness) to robusticity (and roundness) to examine whether or not any demographic patterns exist. Another approach would be to measure the width of the

margin. Buikstra and Ubelaker's (1994) definition for scoring the supraorbital margin recommends palpating the margin, which obviously does not translate to digital observation. That being said, a study that expands on previous research that explored the isolated shape and topography of the supraorbital margin (Graw et al. 1999; Pinto et al. 2006) would be helpful in assessing whether the level of sharpness or roundness can be quantified with volume-rendered CT images and correlated with ordinal scores. The nuchal crest likewise was difficult to score due to the inability to palpate the crest and assess for rugosity and the challenge of depth-perception, so a seriation-type approach would be useful in demonstrating the range of nuchal crest expression as viewed on a computer. Stull et al. (2014) noted that lowering the "opacity ramp", helped with visualizing certain sutures, structures, and overall bone morphology. When a volume-rendered image is constructed, the *opacity curve* determines the "opacity and transparency of various tissues" (Stull et al. 2014: 134). The mental eminence faced the same depth-perception challenges that the nuchal crest did. A future study that viewed a single element under various opacity ramp settings might help with the challenge of depth-perception or a study like Pinto et al. (2016) that uses wavelet analysis or other technology that charts the curves and dips present on a bony element in a two-dimensional format.

In conclusion, the results of this study show that applying population-specific models of sex estimation to populations different than the one the method was developed on produces lower classification rates than either a compatible population-specific model or

a population-inclusive model and should be avoided if possible. The broader impacts for this research indicate that the population-inclusive models will be able to incorporate significantly more variation than population-specific models. While it's unwise to apply models to populations that were not included in the model development, population-inclusive models (with significantly more variation) are better for individuals that may not be included in the model development (i.e., a truly unknown case).

Furthermore, the research of the present study indicates that a population-inclusive method of sex estimation will not produce significantly different accuracy rates compared to population-specific methods of sex estimation. This suggests that a population-inclusive method could be used to estimate the sex of unidentified remains, without knowing the ancestry, with significant accuracy. At the same time, the statistically low accuracy rates that were produced by population-specific models when applied to groups different than the sample used in their development, demonstrate that researchers should be cautious when applying population-specific methods to populations different than the one used to develop the original method.

CHAPTER SIX: CONCLUSION

There has been a recent emphasis on the need for forensic anthropological researchers in the field to suspend their dependence on antiquated methods of ancestry estimation rooted in typological theory and instead focus on understanding the evolutionary mechanisms behind biogeographic differences and the effect that secular change, intersectionality, and the effects of biocultural processes on creating variation across diverse populations (Albanese & Saunders 2006; Bethard & DiGangi 2020; Blakey 2021; Carson 2006a; DiGangi & Bethard 2021).

The results of the present study have shown that population-inclusive nonmetric and metric models of sex estimation will not perform significantly differently than population-specific models, and will perform better, in some cases significantly better than population-specific methods when applied to populations that were not included when the model was developed (e.g., European American model applied to Asian American population). A population-inclusive method of estimating sex from the skull has the potential to supplement currently used ancestry-dependent models of sex estimation (Rogers 2005; Spradley & Jantz 2011; Walker 2008), that generally employ data from either Black or white populations (Tise et al. 2013).

Current ancestry estimations utilized in forensic anthropology are based on assumptions about population variation that are inconsistent with observable patterns of biological variation in different human populations (Albanese & Saunders 2006; Carson 2006a). A population-inclusive method of sex estimation will be an important tool for combatting outdated ancestry-based estimation methods.

Future Directions

The present study only examined the skull, but because the pelvis and long bones offer better success rates for estimating sex (Spradley & Jantz 2011; Patterson & Tallman 2019), it would be fruitful to expand this type of population-inclusive research to include the postcranial skeleton much like the research put forth by Albanese (2008, 2006). The NMDID has a plethora of full-body CT scans that can be utilized for expanding this research to include the postcrania.

Additionally, osteologists suggest that "the most accurate and precise sexing and aging are obtained when it is possible to arrange many skeletal specimens within a series (i.e., seriate) and to compare within in a single biological population" (White 1991: 306). Remains from forensic contexts though, are usually found in isolation and therefore cannot benefit from a seriation-type method and the nonmetric traits are often examined in isolation, without any comparative specimens. For the skull, a study that examines nonmetric traits in complete isolation from the rest of the element and traits (e.g., scoring the mastoid without seeing the glabella) could help elucidate potential observer biases when it comes to morphological sex estimation. Additionally, a large-scale seriation approach could be taken in order to illustrate the wide range of variation.

In addition, the environment, as well as behavior, can influence the development and extent of robusticity of the skeleton which has the potential to affect sex estimation accuracy in more recent populations (Klales 2020; Krogman & Isçan 1986; St. Hoyme & Isçan 1989). Asians living in Asia and Asian-derived populations are already thought to have reduced sexual dimorphism (Tallman 2019; Tallman & Go 2018; Patterson &

Tallman 2019), but it would be useful to look more critically at Asian American populations which may also be affected by increased nutrition. Indeed, this research demonstrated that Asian American classifications are subject to influences from levels of sexual dimorphism that can result in statistically different classification accuracies.

Secular change similarly affects morphological expression of bony traits and the extent of the long-term changes of secular change on sex estimation is currently being examined (Klales 2020). Research indicates that historical collections, like the Terry and Todd collections, are not representative of current forensic anthropological cases due to secular change of the crania and postcrania (Albanese & Saunders 2006; Jantz 2001, Jantz & Meadows Jantz 1999, Spradley & Jantz 2011). In past populations, the levels of observable robusticity, or sexual dimorphism, were greater, particularly in the cranium and mandible (Weiss 1972). This disparity has resulted in biased estimations of fossil specimens being more often classified as male (Weiss 1972). Additionally, over the last two hundred years, the overall size of the cranium has increased slightly (Jantz & Meadows Jantz 2016) and is accompanied by a narrowing of the vault and face resulting in a higher vault and longer basicranium (Jantz 2001; Jantz & Meadows Jantz 2000, 2006; Jantz & Wescott 2002; Wescott & Jantz 2005). The secular changes evidenced in the skeleton are likely attributed to improvements in the health, nutrition, wealth, and a decrease in overall mortality and morbidity rate in humans (Klales 2020; Kilroy et al. 2020). The present study could be expanded to include historical and ancient population-diverse samples in order to assess the effect of secular change on a population-inclusive model.

Previous research has demonstrated that CT scans can be used to gather precise data as well as accurate estimates of sex (Decker et al. 2011; Herrera & Tallman 2019; Richard et al. 2014; Robinson et al. 2008; Stull et al. 2014). This study similarly produced classification results in line with hands-on sex estimation studies. It would be useful to continue to explore CT scans and volume-rendered imaging as a potential replacement method for hands-on skeletal analysis, especially in cases where soft tissue cannot be removed.

Limitations of the Current Study

One of the biggest limitations of the present study was establishing an equally distributed sample across sex, population affiliation, and age. This was in part due to the availability of the NMDID, which is disproportionately comprised of white males, but also to issues with downloading the large files of CT images for a large sample size. In addition to the demographic limitations, another obstacle was the completeness of the CT scans which were variably affected by the cause of death. This incompleteness was often not readily apparent until after the CT-image had been volume rendered and debris removed. On the other hand, this type of sample is directly representative of the type of remains found in forensic contexts (i.e., incomplete).

The volume-rendering process was also subject to PVE, as previously discussed, which limited data collection in some cases. Although part of the sample preparation included removing artifacts that obstructed craniometrics landmarks or morphological traits, some PVE could be removed without removing part of the skull as well.

Another limitation was that the nonmetric scoring of the nuchal crest and supraorbital margin was challenging because of the quality of the CT image. Establishing depth and estimating the topography of both traits was difficult which is likely why both traits had the lowest level of intraobserver agreement.

In conclusion, this research indicates that population-inclusive methods of sex estimation should be further explored, developed, validated, and become part of forensic anthropologists' standard toolkit. While population-specific methods of identification have a clear utility, the scope of populations that have validated estimation methods is lacking, which necessitates the use of population-inclusive methods. Furthermore, just because European American population-specific method *can* be applied to non-European populations with poor-moderate success, does not mean it *should* be. The broader implications of such practices need to be carefully considered and researchers should use caution when perpetuating these types of practices, especially in light of recent discussion regarding the role of ancestry in the biological profile (Adams & Pilloud 2020; Albanese and Saunders 2006; Bethard & Digangi 2020; Blakey 2021; Carson 2006a; DiGangi and Bethard 2021; Edgar 2020; Ross et al. 2020; Spradley & Jantz 2020; Tallman et al. 2021).

BIBLIOGRAPHY

Acsádi G and Nemeskéri J. 1970. History of human life span and mortality. Budapest: Akadémiai Kiadó.

Adams DM and Pilloud MA. 2020. Ancestry estimation, race science, and scientific racism in forensic anthropology. In: Proceedings of the 72nd Annual Scientific Meeting of the American Academy of Forensic Sciences; 2020 February 17-22; Anaheim, CA. Colorado Springs, CO: American Academy of Forensic Sciences;223.

Albanese J and Saunders SR. 2006. Is it possible to escape racial typology in forensic identification? In A. Schmitt, E Cunha, and J Pinheiro (Eds.). *Forensic Anthropology and Medicine. Complementary Sciences from Recovery to Cause of Death*, Humana Press, p. 281-216.

Albanese J, Eklics G, and Tuck A. 2008. A metric method for sex determination using the proximal femur and fragmentary hipbone. *Journal of Forensic Science,* 53(6): 1283-8.

Albanese J, Tuck A, Gomes J, and Cardoso HFV. 2016. An alternative approach for estimating stature from long bones that is not population-or group-specific. *Forensic Science International,* 259:59-68.

Arsuaga JL and JM Carretero. 1994. Multivariate analysis of the sexual dimorphism of the hip bone in modern human population and in early hominids. *American Journal of Physical Anthropology*, 93:241-257.

Austin D and King RE. 2016. The biological profile of unidentified human remains in a forensic context. *Academy of Forensic Pathology,* 6(3):370-390.

Badyaev AV. 2002. Growing apart: an ontogenetic perspective on the evolution of sexual size dimorphism. *Trends in Ecology and Evolution,* 17:369-378.

Bass WM. 1995. Human osteology: a laboratory and field manual. Columbia Missouri Archaeological Society.

Bass WM. 2005. Human osteology: a laboratory and field manual, 5th edition. Columbia, MO: Missouri Archaeological Society.

Baughan B and Demirjian A. 1978. Sexual dimorphism in the growth of the cranium. *American Journal of Physical Anthropology,* 49:383-90.

Beckmann EC. 2006. CT scanning in the early days. *British Journal of Radiology*, 79(937):5-8.

Benya EC. 2008. Advances in computed tomography. *Pediatric Analysis,* 37(6):428-431.

Berstein RM. 2010. The big and small of it: how body size evolves. *American Journal of Physical Anthropology,* 143 (suppl. 51): 46-62.

Berry S R. 2014. Metadata determination for cadaveric collection. Thesis, University of New Mexico, Albuquerque, New Mexico.

Best KC, Garvin MS and Cabo LL. 2018. An investigation into the relationship between human cranial and pelvic sexual dimorphism. *Journal of Forensic Sciences*, 63(4):990-1000.

Bethard JD and DiGangi EA. 2020. Moving beyond a lost cause: forensic anthropology and ancestry estimates in the United States. *Journal of Forensic Sciences,* 65(5):1791-1792.

Blackless M, Charuvastra A, Derryck A, Fausto-Sterling A, Lauznne K, and Lee E. 2000. How sexually dimorphic are we? Review and synthesis. *American Journal of Human Biology,* 12(2):151-166.

Blakey ML. 1987. Skull doctors: Intrinsic social and political bias in the history of American physical anthropology with special reference to the work of Aleš Hrdlicˇka. *Critique of Anthropology*, 7(2): 7–35.

Blakey ML. 2021. Understanding racism in physical (biological) anthropology. *American Journal of Physical Anthropology,* 175(2):316-325.

Blau S, Robertson S, and Johnston M. 2008. Disaster victim identification: new applications for post mortem computed tomography. *Journal of Forensic Sciences,* 53(4):956-961.

Bouillon R, Bex M, Vanderschueren D, and Boonen S. Estrogens are essential for male pubertal periosteal bone expansion. *Journal of Clinical Endocrinology & Metabolism*, 89(12):6025-9.

Broca P. 1875. Instructions craniologiques et craniométriques. *Society of Anthropology of Paris,* TII. Series 2.

Brough A, Morgan B, and Rutty G. 2015. Postmortem computed tomography (PMCT) and disaster victim identification. *La Radiologia Medica,* 120(9):866-873.

Buikstra JE and Ubelaker DH, editors. 1994. Standards for data collection from human skeletal remains. Fayetteville: Arkansas Archaeological Survey, Research Series No. 44.

Burns K. 1999. Forensic anthropology training manual. New Jersey: Prentice Hall.

Burns KR. 2006. Forensic anthropology training manual (2nd ed.). New York: Pearson

Byers S. 2002. Introduction to forensic anthropology. Boson: Allyn and Bacon.

Camargo JR, Daruge E, Prado FB, Caria PHF, Alves MC, Silva RF, and Dargue Jr. E. 2007. *Brazilian Journal of Morphological Science*, 24(4):239-243.

Carson EA. 2006a. Maximum likelihood estimation of human craniometrics heritabilities. *American Journal of Physical Anthropology,* 131(2):169-80.

Carson EA. 2006b. Maximum-likelihood variance components analysis of heritabilities. *Human Biology*, 78(4):383-402.

Carson JA and Manolagas SC. 2015. Effects of sex steroids on bones and muscles: similarities, parallels, and putative interactions in health and disease. *Bone,* 80:67-78.

Christensen AM and Crowder CM. 2009. Evidentiary standards for forensic anthropology. *Journal of Forensic Sciences,* 54(6):1211-1216.

Cinamon U. 2009. The growth rate and size of the mastoid air cell system and mastoid bone: a review and reference. *European Archivers of Oto-Rhino-Laryngology,* 266: 781-786.

Cohen J. 1968. Weighted Kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *American Psychological Association, Inc.,* 70(4):213-20.

Cohen J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurements*, 20(1):37-46.

Cunha E and van Vark G. 1991. The construction of sex discriminant functions from a large collection of skulls of known sex. *International Journal of Anthropology*, 6:53-66.

Curtis AE. 2017. Sex and age at death estimation from the os pubis: validation of two methods on a modern autopsy sample. Thesis, Boston University, Boston, MA.

da Silva RF, Prado FB, Caputo IGC, Devito KL, de Luscena Botelho T, and Daruge Jr E. 2009. The forensic importance of frontal sinus radiographs. *Journal of Forensic and Legal Medicine,* 16:18-23.

Dayal MR, Spocter MA, and Bidmos MA. 2008. An assessment of sex using the skull of black South Africans by discriminant function analysis. *Journal of Comparative Human Biology,* 59:209-221.

Daneshvari Berry S and Edgar HJH. Development of a large-scale, whole body CT image database. AMIA Annual Symposium Proceedings. 2017:1951.

Daneshvari Berry S and Edgar HJH. 2021. Announcement: the New Mexico decedent image database. *Forensic Imaging,* 24.

Decker SJ, Davy-Jow SL, Ford JM, and Hilbelink DR. 2011. Virtual determination of sex: metric and nonmetric traits of the adult pelvis from 3D computed tomography models. *Journal of Forensic Sciences,*56(5).

DiGangi EA and Bethard JD. 2021. Uncloaking a lost cause: decolonizing ancestry estimation in the United States. *American Journal of Physical Anthropology,* 175:422-436.

DuBois LZ and Shattuck-Heidorn H. 2021. Challenging the binary: gender/sex and the bio-logics of normalcy. *American Journal of Human Biology*, 33:e23623.

Edgar HJH. 2020. Population structure, population, heterogeneity, and sources of error in the forensic estimation of "race". In: Proceedings of the 72nd Annual Scientific Meeting of the American Academy of Forensic Sciences; 2020 February 17-22; Anaheim, CA. Colorado Springs, CO: American Academy of Forensic Sciences;216.

Edgar HJH, Danshvari Berry S, Moes E, Adolphi NL, Bridges P, and Nolte KB. 2020. New Mexico Decedent Image Database. Office of the Medical Investigator, University of New Mexico.

Edmond G and Roach K. 2011. A contextual approach to the admissibility of the state's forensic science and medical evidence. *The University of Toronto Law Journal*, 61(3):343-409.

Eleveth PB and Tanner JM. 1990. Worldwide variation in human growth (2nd ed.) Cambridge: Cambridge University Press.

Endo B. 1970. Analysis of stresses around the orbit due to masseter and temporalis muscles. *Journal of the Anthropological Society of Nippon,* 78:251-266.

Ferembach D. Schwidetzky I, and Stloukal M. 1980. Recommendations for age and sex diagnoses of skeletons. *Journal of Human Evolution,* 9:517-549.

Field A. 2009. Discovering statistics using SPSS. 3rd ed. Thousand Oaks CA: SAGE.

France DL. Observation and metric analysis of sex in the skeleton. In: Reichs KJ, editor. Forensic osteology: advances in the identification of human remains. Springfield, IL: Charles C. Thomas, 163-86.

Franklin D, Freedman L, and Milne N. 2005. Sexual dimorphism and discriminant function sexing in indigenous South African crania. *HOMO—Journal of Comparative Human Biology,* 55(3):213-228.

Frayer D and Wolpoff M. 1985. Sexual dimorphism, *Annual Review of Anthropology*, 14:429-473.

Freidline SE, Gunz P, and Hublin JJ. 2015. Ontogentic and static allometry in the human face: contrasting Khoisan and Inuit. *American Journal of Physical Anthropology,* 158(1):116-131.

Gaballah IF, Shehab AM, and Bayoumi KA. 2014. Sex determination in femurs of modern Egyptians: a comparative study between metric measurements and SRY gene detection. *Egyptian Journal of Forensic Sciences*, 4(4):109-115.

Gardner J, Garvin G, Nelson A, Vascotto G, and Conlogue G. 2004. Paleoradiology in mummy studies: the Sulman mummy project. *Canadian Association of Radiologists Journal*, 55(4):228-234.

Garvin HM. 2012. Adult sex determination: methods and application. In Dirkmaat DC, (eds.) A companion to forensic anthropology. Malden, MA: Blackwell Publishing Ltd.

Garvin HM and Ruff CB. 2012. Sexual dimorphism in skeletal browridge and chin morphologies determined using a new quantitative method. *American Journal of Physical Anthropology,* 147:661-70.

Garvin HM, Sholts SB, and Mosca LA. 2014. Sexual dimorphism in human cranial trait scores: effects of population, age, and body size. *American Journal of Physical Anthropology*, 154: 259-269.

GLAAD. 2021. https://www.glaad.org/reference/transgender. Accessed September 16th 2021.

Graw M, Czarnetzki A, and Haffner HT. 1999. The form of the supraorbital margin as a criterion in identification of sex from the skull: investigations based on modern human skulls. *American Journal of Physical Anthropology*, 108:91-96.

Gray H. 1966. Gray's Anatomy. Edited by CM Goss. Lea and Febiger, Philadelphia, PA.

Grivas CR and Komar DA. 2008. Kumho, Daubert, and the nature of scientific inquiry: implications for forensic anthropology. *Journal of Forensic Sciences,* 53(4):771-776.

Hefner JT. 2009. Cranial nonmetric variation and estimating ancestry. *Journal of Forensic Science,* 54(5):985-995.

Herrera MD and Tallman SD. 2019. Craniometric variation and ancestry estimation in two contemporary Caribbean populations. *Forensic Science International*, 305.

Hilloowala RA and Trent RB. 1988. Supraorbital ridge and masticatory apparatus II: humans (Eskimos). *Human Evolution,* 3:351-356.

Hrdlicka A. 1952, Practical anthropometry. 4th edition edited by TD Stewart. Wistar Institute of Anatomy and Biology, Philadelphia, PA.

Hughes CE, Tise ML, Trammell LH, and Anderson BE. 2013. Cranial morphological variation among contemporary Mexicans: regional trends, ancestral affinities, and genetic comparisons. *American Journal of Physical Anthropology,* 151:506-517.

Humphrey LT. 1998. Growth patterns in the modern human skeleton. *American Journal of Physical Anthropology,* 105:57-72.

Hunley K, Edgar H, Healy M, Mosley C, Cabana GS, and West F. 2017. Social identity in New Mexicans of Spanish-speaking descent highlights limitations of using standardized ethnic terminology in research. *Human Biology,* 89(3):217-228.

Işcan MY. 2005. Forensic anthropology of sex and body size. *Forensic Science International,* 147(2-3):107-112.

Jantz RL. 2001. Cranial change in Americans: 1850-1975. *Journal of Forensic Sciences,* 46(4):784-7.

Jantz RL and LM Jantz. 1999. Secular change in long bone length and proportion in the United States, 1800-1970. *American Journal of Physical Anthropology,* 111:57-67.

Jantz RL and L Meadows Jantz. 2000. Secular change in craniofacial morphology. *American Journal of Human Biology*, 12:327-338.

Jantz RL and L Meadows Jantz. 2016. The remarkable change in Euro-American cranial shape and size. *Human Biology,* 88(1):56-64.

Jantz R and Ousley S. 2013. Fordisc 3: third generation of computer-aided forensic anthropology. *Rechtsmedizin*, 23:97-99.

Jantz RL and Wescott DJ. 2002. Assessing craniofacial secular change in American whites and blacks using geometric morphometry. *American Journal of Physical Anthropology,* (Suppl.)34:90.

Jones R. 2020, April 16th. NMDID aids in COVID-19 research. *The University of New Mexico.* Retrieved from https://hsc.unm.edu/news/news/nmdid-aids-in-covid-19-research.html.

Karsten JK. 2017. A test of the preauricular sulcus as an indicator of sex. *American Journal of Physical Anthropology*. 1-5.

Kemkes A and Gobel T. 2006. Metric assessment of the "mastoid triangle" for sex determination: a validation study. *Journal of Forensic Sciences,* 51:985-989.

Kincer C and Tallman S. 2019. Transgender knowledge in forensic anthropology survey. Qualtrics, Boston University School of Medicine.

Klales AR. 2013. Current practices in physical anthropology for sex estimation in unidentified, adult individuals. *American Journal of Physical Anthropology,* 150(56):168.

Klales AR. 2016. Secular change in morphological pelvic traits used for sex estimation. *Journal of Forensic Sciences*, 61(2).

Klales AR. 2018. MorphoPASSE: the morphological pelvis and skull sex estimation database. Version 1.0. Washburn University, Topeka, KS.

Klales AR. 2020. Sex estimation of the human skeleton: history, methods, and emerging techniques. Waltham: Elsevier Press.

Klales AR and Cole SJ. 2017. Improving nonmetric sex classification for Hispanic individuals. *Journal of Forensic Science,* 62(4):975-980.

Klales AR, Ousley SD, and Vollner JM. 2012. A revised method of sexing the human innominate using Phenice's nonmetric traits and statistical methods. *American Journal of Physical Anthropology,* 149:104-114.

Krogman WM. 1955. The skeleton in the forensic medicine. *Postgraduate Medical Journal,* 17(2): A48-A62.

Krogman WM. 1962. The human skeleton in forensic medicine. Charles C. Thomas, Springfield IL.

Krogman WM and Işcan MY. 1986. The human skeleton in forensic medicine. Springfield, IL: Charles C. Thomas.

Landis JR and Koch GG. 1977. The measurement of observer agreement for categorical data. *Biometrics,* 33(1):159-174.

Langley NR, Meadows Jantz L, Ousley SD, Jantz RL, and Milner G. 2016. *Data collection procedures for forensic skeletal material 2.0*. Knoxville, TN: Forensic Anthropology Center.

Lahr MM and Wright RVS. 1996 The question of robusticity and the relationship between cranial size and shape in *Homo sapiens*. *Journal of Human Evolution,* 31:157-191.

Letterman GS. 1941. The greater sciatic notch in American Whites and Negroes. *American Journal of Physical Anthropology,* 98:59-72.

Lewis CJ and Garvin HM. 2016. Reliability of the Walker cranial nonmetric method of implications for sex estimation. *Journal of Forensic Sciences,* 61(3):743-751.

Lieberman DE. 2011. The evolution of the human head. Cambridge: Harvard University Press.

Lieberman DE, Hallgrimsson B, Liu W, Parsons TW, and Jamnickzky HA. 2008. Spatial packing, cranial base angulation, and craniofacial shape variation in the mammalian skull: testing a new model using mice. *Journal of Anatomy,* 212:720-735.

Lovell NC. 1989. Test of Phenice's technique for determining sex from the os pubis. *American Journal of Physical Anthropology,* 79:117-120.

Martin R and Knussman R. 1988. Anthropologie: Handbuch Der Vergleichenden Biologie Des Menschen. Gustav Fisher Verlag, Stuttgart.

Mendéndez LP. 2017. Comparing methods to assess intraobserver measurement of error of 3D craniofacial landmarks using geometric morphometrics through a digitizer arm, *Journal of forensic Science,* 62(3):741-746.

Montagu MFA. 1960. Introduction to physical anthropology. Charles C Thomas, Springfield, IL.

Moss JL. 2018. The forgotten victims of missing white woman syndrome: an examination of legal measures that contribute to the lack of search and recovery of missing black girls and women. *William & Mary Journal of Race, Gender, and Social Justice,* 25:737-62.

Moss ML and Young RW. 1960. A functional approach to craniology. *American Journal of Physical Anthropology,* 18:281-292.

National Institute of Biomedical Imaging and Bioengineering (NIBIB). 2021. https://www.ni bib.nih.gov/science-education/science-topics/computed-tomography-ct. Accessed September 25th, 2021.

Noren A, Lynnerup N, Czarnetzki A and Graw M. 2005. Lateral angle: a method for sexing using the petrous bone. *American Journal of Physical Anthropology,*128:318–323.

Osipov B, Havarti K, Nathena D, Spanakis K, Karantanas A, and Kranioti EF. 2013. Sexual dimorphism of the bony labyrinth: A new age-independent method. *American Journal of Physical Anthropology,* 151:290-301.

Patterson MM and Tallman SD. 2018. Cranial and postcranial metric sex estimation in modern Thai and Ancient Native American individuals. *Forensic Anthropology,* 2(4):1-20.

Petarosa A, Sholts SB, Slaus M, Bosnar A, and Wärmländer SKTS. 2015. Evaluating sexual dimorphism in the human mastoid process: a viewpoint on the methodology. *Clinical Anatomy,* 28:593-601.

Petrik V, Apok V, Britton JA, Bell BA, and Papadopoulos MC. 2006. Godfrey Hounsfield and the dawn of computed tomography. *Neurosurgery,* 58(4):780-787.

Phenice TW. 1969. A newly developed visual method of sexing the os pubis. *American Journal of Physical Anthropology.* 30(20:297-301.

Pickering RB and Bachman DC. 1997. The use of forensic anthropology. Boca Raton, FL: CRC Press.

Pinto SC, Urbanová P, and Ceaser RM Jr. 2016. Two-dimensional wavelet analysis of supraorbital margins of the human skull for characterizing sexual dimorphism. *IEEE Transactions on Information Forensics and Security,* 11(7):1542-1548.

Relethford JH. 1994. Craniometric variation among modern human populations. *American Journal of Physical Anthropology,* 95(1):53-62.

Rennie SR. 2018. Summary sex: a multivariate approach to sex estimation from the human pelvis. Unpublished Doctoral Dissertation. Liverpool John Moores University, Liverpool, England, United Kingdom.

Richard AH, Parks CL, and Monson KL. 2014. Accuracy of standard craniometrics measurements using multiple data formats. *Forensic Science International*, 242:177-185.

Roberts CA. 1996. Forensic anthropology 1: the contribution of biological anthropology to forensic contexts. In: Hunter J, Roberts C, Martin A, (eds). Studies in crime: an introduction to forensic archaeology. London: Routledge, p. 101-21.

Robinson C, Eisma R, Morgan B, Jeffery A, Graham EAM, and Black S.

Rogers TL. 1999. A visual method of determining the sex of the skeleton using the distal humerus. *Journal of Forensic Sciences,* 44(1):57-60.

Rogers TL. 2005 Determining the sex of human remains through cranial morphology. *Journal of Forensic Sciences,* 50(3):1-8.

Rogers TL and Saunders S. 1994. Accuracy of sex determination using morphological traits of the human pelvis. *Journal of Forensic Sciences,* 39:1047-1056.

Ross AH. 2004. Cranial evidence of pre-contact multiple population expansions in the Caribbean. *Caribbean Journal of Science,* 40(3):291-298.

Ross AH, Bethard JD, DiGangi EA. 2020. A call for a paradigm shift in the study of ancestry. In: Proceedings of the 72nd annual Scientific Meeting of the American Academy of Forensic Sciences; 2020 February 17-22. Anaheim, CA. Colorado Springs, CO. American Academy of Forensic Sciences, 222.

Roth M. Ousley SD, Tuamsuk P. 2013. Sex estimation using non-metric traits in Thai crania with the Walker (2008) method. In: Proceedings of the 65th Annual Meeting of the American Academy of Forensic Sciences, February 18-23. Washington, DC.

Russell MD. 1985. The supraorbital torus: a most remarkable peculiarity. *Current Anthropology,* 26:337-360.

Saggese G, Baroncelli GI, Bertelloni S. 2002. Puberty and bone development. *Best Practice & Research: Clinical Endocrinology & Metabolism*, 16(1):53-64.

Saini V, Srivastava R, Rai RK, Shamal SN, Singh TB, and Tripathi SK. 2012. Sex estimation from the mastoid process among north Indians. *Journal of Forensic Science,* 57:434-439.

Salinas C Jr. 2020. The complexity of the "x" in Latinx: how Latinx/a/o students relate to, identify with, and understand the term Latinx. *Journal of Hispanic Higher Education*, 19(2):149-168.

Sardi ML, Joosten GG, Pandiani CD, Gould MM, Anzelmo M, and Ventrice F. 2018. Frontal sinus ontogeny and covariation with bone structures in a modern human population. *Journal of Morphology,* 279(7):871-882.

Scheuer L and Black S. 2004. The Juvenile Skeleton USA: Academic Press.

Schultz AH. 1930. The skeleton of the trunk and limbs of higher primates. *Human Biology*, 2:303-438.

Sommers Z. 2016. Missing white woman syndrome: an empirical analysis of race and gender disparities in online news coverage of missing persons. *Journal of Criminal Law & Criminology*, 106(2):275-314.

Spradley MK. 2016. Metric methods for the biological profile in forensic anthropology: sex, ancestry, and stature. *Academy of Forensic Pathology*, 6(3): 391-399.

Spradley MK and Jantz RL. 2011. Sex estimation in forensic anthropology: skull versus postcranial elements. *Journal of Forensic Sciences*. 56(2): 289-296.

Spradley KS and Jantz R. 2020. What are we really estimating? An evaluation of what we commonly refer to as ancestry estimation. In: Proceedings of the 72nd Annual Scientific Meeting of the American Academy of Forensic Sciences; 2020 February 17-22; Anaheim, CA. Colorado Springs, CO: American Academy of Forensic Sciences;217

Spradley MK, Jantz RL, Robinson A, Peccerelli F. 2008. Demographic change and forensic identification: problems in metric identification of Hispanic skeletons. *Journal of Forensic Sciences,* 53(1):21-8.

St. Hoyme LE and Işcan MY. 1989. Determination of sex and race: accuracy and assumptions. In MY Işcan and KAR Kennedy (eds.): Reconstruction of life from the skeleton. New York:Alan R Liss, p 53-93.

Standring S. 2008. Gray's Anatomy. 40th Ed. Spain: Churchill Livingstone.

Stewart TD. 1954. Evaluation of evidence form the skeleton. In Legal Medicine by RBH Gradwohl. CV Mosley, St. Louis, MO.

Stewart T. 1979. Essentials for forensic anthropology. Springfield, IL: Charles C. Thomas.

Stinson S. 1985. Sex differences in environmental sensitivity during growth and development. *Yearbook of Physical Anthropology*, 28:123-147.

Stinson S. 2012. Growth variation: biological and cultural factors. In: Stinson S, Bogin B, O'Rourke D, (eds.) Human biology: an evolutionary and biocultural perspective, 2nd edition. Hoboken: Wiley. P.587-635.

Stock MK. 2018. A preliminary analysis of the age of full expression of sexually dimorphic cranial traits. *Journal of Forensic Sciences,* 63(6):1802-1808.

Stull KE, Tise ML, Ali Z, and Fowler DR. 2014. Accuracy and reliability of measurements obtained from computed tomography 3D-volume rendered images. *Forensic Science International*, 238:133-140.

Sutherland LD and Suchey JM. 1991. Use of ventral arc in pubic sex determination. *Journal of Forensic Sciences*, 36(2):501-511.

Sutphin R, Ross AH, Jantz RL. 2014. Secular trends in Cherokee cranial morphology: eastern and western bands. *Annals of Human Biology*, 41(6): 511-517.

Tallman SD. 2019. Cranial nonmetric sexual dimorphism and sex estimation in east and southeast Asian individuals. *Forensic Anthropology*, 2(4):1-18.

Tallman SD and Go MC. 2018. Application of the optimized summed scored attributes method to sex estimation in Asian crania. *Journal of Forensic Sciences,* 63(3):809-814.

Tallman SD, Kincer CD, and Plemons ED. 2021. Centering transgender individuals in forensic anthropology and expanding binary sex estimation in casework and research.

Tise ML, Spradley MK, and Anderson BE. 2013. Postcranial sex estimation of individuals considered Hispanic. *Journal of Forensic Sciences,* 58(S1):S9-S14.

Ubelaker DH and DeGaglia CM. 2017. Population variation in skeletal sexual dimorphism. *Forensic Science International,* 278:407.e1-407.e7.

Vance VL, Steyn M, and L'Abbé ENL. 2011. Nonmetric sex determination from the distal and posterior humerus in Black and white South Africans, *Journal of Forensic Science,* 56(3): 710-714.

Vanderschueren D, Vandenput L, Boonen S, Lindberg ML, Bouillon R, and Ohlsson C. 2004. Androgens and bone. *Endocrine Reviews,* 25(3):389-425.

Vanderschueren D, Venken K, Ophoff J, Boullon R, Boonen S. 2006. Sex steroids and the periosteum-reconsidering the roles of androgens and estrogens in periosteal expansion. *Journal of Clinical Endocrinology & Metabolism,* 91(2):378-82.

Viðarsdóttir US, O'Higgins P, and Stringer C. 2002. A geometric morphometric study of regional differences in the ontogeny of modern human facial skeleton. *Journal of Anatomy*, 201(3):211-229.

Vinyard CJ and Smith FH. 1997. Morphometric relationships between the supraorbital region and frontal sinus in Melanesian crania. *Journal of Comparative Human Biology,* 48: 1-21.

Vinyard CJ and Smith FH. 2001. Morphometric testing of structural hypotheses of the supraorbital region in modern humans. *Zeitschrift für Morphologie und Anthropologie*, 83:23-41.

Vollner JM, Klales AR, Ousley SD. 2009. Sexing of the human pelvis using non-metric traits and statistical analysis. Proceedings of the 78th Meeting of the American

Association for Physical Anthropology, March 31st to April 4th; Chicago, IL. *American Journal of Physical Anthropology*, 138(48 suppl):263-264.

Wahl J and Graw M. 2001. Metric sex differentiation of the pars petrosa ossis temporalis. *International Journal of Legal Medicine,* 114:215–223.

Waldron T. 1987. The relative survival of the human skeleton: implications for paleopathology. In AN Boddington and RC Janaway (Eds.), *Death, Decay, and Reconstruction,* Manchester University Press, Manchester, p. 55-64.

Walker PL. 1995. Problems of preservation and sexism in sexing: some lessons from historical collections for palaeodemographers. In: Saunders SR, Herring A (Eds.) *Grave Reflections, Portraying the Past through Cemetery Studies*. Toronto: Canadian Schoalrs' Press. P. 31-47.

Walker PL. 2005. Greater sciatic notch morphology: sex, age, and population differences. American Journal of Physical Anthropology, 127:385-391.

Walker PL. 2008. Sexing skulls using discriminant function analysis of visually assessed traits. *American Journal of Physical Anthropology*, 136:39-50.

Walrath DE, Turner P and Bruzek J. 2004. Reliability test of the visual assessment of cranial traits for sex determination. *American Journal of Physical Anthropology,* 125:132–137.

Washburn SL. 1948. Sex difference in the pubic bone. *American Journal of Physical Anthropology*, 6:199-207.

White T. 2002. Human osteology (2nd edition). New York: Academic Press.

White TD, Black MT, and Folkens PA. 2012. Human Osteology (3rd edition). Amsterdam: Elsevier Press.

Williams BA, Rogers T. 2006. Evaluating the accuracy and precision of cranial morphological traits for sex determination. *Journal of Forensic Science,* 51:729–735.

Weiss KM. 1972. On the systematic bias in skeletal sexing. *American Journal of Physical Anthropology,* 37(2):239-49.

Weidenreich F. 1936. The mandibles of *Sinathropus pekinensis:* a comparative study. *Palaeontologia Sinica. Series D*, 7:1-62.

Weidenreich F. 1941. The brain and its role in the phylogenetic transformation of the human skull. *Transactions of the America Philosophical Society,* 31:321-442.

Wescott DJ and Jantz RL. 2005. Assessing craniofacial secular change in American blacks and whites using geometric morphometry. In DE slice (ed.), Modern morphometrics in physical anthropology, New York: Kluwer Academic/Plenum, p.231-245.

World Health Organization. Gender and Health. https://www.who.int/news-room/q-a-detail/gender-and-health. Accessed September 16th 2021.

CURRICULUM VITAE