Testing Population-Inclusive Assigned Sex at Birth Estimation Methods of the Skull

Victoria Conquest[1] and Sean D. Tallman[1,2]

[1] Department of Anatomy & Neurobiology, Boston University Chobanian & Avedisian School of Medicine

[2] Department of Anthropology, Boston University

Introduction

Historically, assigned sex at birth (ASAB) estimation techniques have relied on ancestry estimation and population-specific methods (Dayal et al. 2008; Garvin et al. 2014; Rogers 2005). Recently, ancestry or population affiliation estimation has been questioned due to the equivocal nature of many forensic ancestry traits and the dangers associated with the reification of biological race, among other potential harms (Bethard and DiGangi 2020; DiGangi and Bethard 2021). In cases where ancestry is intentionally excluded from the biological profile, as some forensic anthropologists are doing, or in truly unknown forensic cases, population-inclusive ASAB estimation methods incorporating significant variation are vital. Kelley and Tallman (2022) developed population-inclusive nonmetric, metric, and combined nonmetric/metric ASAB estimation methods on a demographically diverse U.S. sample of computed tomography (CT) scans of the skull (n=431), including African American (m=40, f=43), Asian American (m=69, f=20), European American (m=43, f=41), Latin American (m=44, f=43), and Native American (m=46, f=42) individuals. Binary logistic regressions (BLR) and discriminant function analyses (DFA) were employed in order to develop population-specific and population-inclusive models. Kelley and Tallman (2022) found that population-inclusive nonmetric and metric models of sex estimation did not perform significantly differently than population-specific models for most groups, with correct classifications ranging from 81.0-91.6%. The current research tests Kelley and Tallman's (2022) population-inclusive CT-derived methods on skeletonized individuals from the University of Tennessee, Knoxville's (UTK) Donated Skeletal Collection.

Materials and Methods

Standard cranial nonmetric (Walker 2008) and metric (Spradley and Jantz 2011) variables were collected on 439 adult individuals (18-90 years), including 217 assigned females at birth (AFAB) and 222 assigned males at birth (AMAB), from UTK's Donated Skeletal Collection. Though randomized, the majority (98.8%) of the individuals were European American due to the homogenous demographics of UTK's Donated Skeletal Collection. Additionally, a colleague familiar with the methods scored 10.3% (n=45) of the sample and interobserver agreement was calculated using intraclass correlation coefficients (ICC) and Cohen's kappa statistics.

Five nonmetric traits were scored using the methods outlined by Walker (2008) and included the glabella, supraorbital margin, nuchal crest, mastoid process, and mental eminence. Additionally, 18 standard points of cranial and mandibular measurements were collected following the protocols outlined by Buikstra and Ubelaker (1994) and Spradley and Jantz (2011).

Kelley and Tallman's CT-derived population-inclusive nonmetric, metric, and combined models were applied to the skeletonized individuals. Furthermore, UTK-specific BLM and DFA models, including the same combination of traits as Kelley and Tallman's (2022) models (UTK-KT), were developed in IBM SPSS (version 29) to compare population-specific and population-inclusive methods. Model equations are presented in Table 1.

Results

Classification accuracies for Kelley and Tallman (2022), UTK-KT, and UTK-specific models are presented in Table 2. The original Kelley and Tallman (2022) methods performed with the lowest accuracies with only the combined model correctly classifying 82.9%. The classification accuracies of AFAB individuals performed particularly low with the nonmetric and combined models performing at 78.8% and 71.4%, respectively. The classification accuracy for AMAB individuals in the metric model performed at 78.4%. The UTK-specific models performed better than Kelley and Tallman's (2022) models and the UTK-KT models(nonmetric 89.5%, metric 86.7%, and combined 90.2%).

The ICC analysis found that of the 18 measurements, four of them had poor reliability (upper facial height, nasal height, foramen magnum breadth, and mastoid length), one had good reliability (occipital chord), and the other thirteen had excellent reliability (Koo and Li 2016). The Cohen's kappa analysis found that the nuchal crest, supraorbital margin, and the mental eminence had a slight agreement and the mastoid process and the glabella had fair agreement (Landis and Koch 1978). These somewhat low agreement rates may be attributed to differences in the interpretation or understanding of the measurements between observers.

Table 1. Classificatory BLR and DFA equations used in this study.

Method	Equations
Kelley and Tallman (2022) population inclusive nonmetric model	(glabella score*1.385) + (mastoid score*0.902) + (mental eminence score*0.44) + (-5.888) (above 0.5 suggests male)
Kelley and Tallman (2022) population inclusive metric model	(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) (above -0.221 suggests male)
Kelley and Tallman (2022) population inclusive combined model	(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) (above 0.5 suggests male)
UTK-KT nonmetric model	(glabella score*2.004) + (mastoid score*1.132 + (mental eminence score*0.638) + (-9.744) (above 0.5 suggests male)
UTK-KT metric model	(glabella occipital length*0.031) + (bizygomatic breadth*0.125) + (biauricular breadth*0) + (minimum frontal breadth*-0.041) + (nasal height*-0.001) + (orbital height*0.005) + (mastoid height*0.068) + (bigonial breadth*0.022) + (maximum ramus height*0.089) + (mandibular length*0.006) + (-27.855) (above -0.0215 suggests male)
UTK-KT combined model	(glabella score*2.001) + (mastoid score*0.654) + (mental eminence score*0.521) + (glabella occipital length*0.041) + (bizygomatic breadth*0.265) + (maximum ramus height*0.188) + (mandibular length*0.049) + (-64.849) (above 0.5 suggests male)
UTK-specific nonmetric model	(nuchal crest score*1.113)+(mastoid score*1.119)+(glabella score*1.667)+(mental eminence score*0.646)+(-11.895) (above 0.5 suggests male)
UTK-specific metric model	(glabella occipital length*0.020)+(bizygomatic breadth*0.132)+(minimum frontal breadth*-0.038)+(frontal chord*0.031)+(mastoid length*0.075)+(maximum ramus height*0.085)+(foramen magnum breadth*0.057)+(-29.224) (above -0.0325 suggests male)
UTK-specific combined model	(nuchal crest score*1.257)+(glabella score*1.958)+(bizygomatic breadth*0.303)+(mastoid length*0.241)+(maximum ramus height*0.261)+(-69.646) (above 0.5 suggests male)

Table 2. Classification accuracies of models.

Method	Combined Classification Accuracy (n=439)	AFAB Classification Accuracy (n=217)	AMAB Classification Accuracy (n=222)
Kelley and Tallman (2022) population inclusive nonmetric model	85.6%	78.8%	92.3%
Kelley and Tallman (2022) population inclusive metric model	79.3%	80.2%	78.4%
Kelley and Tallman (2022) population inclusive combined model	82.9%	71.4%	94.1%
UTK-KT nonmetric model	87.9%	94.0%	82.0%
UTK-KT metric model	85.6%	83.8%	87.4%
UTK-KT combined model	89.1%	90.8%	87.4%
UTK-specific nonmetric model	89.5%	93.1%	86.0%
UTK-specific metric model	86.7%	85.3%	87.8%
UTK-specific combined model	90.2%	91.2%	89.2%

Discussion and Conclusions

The nonmetric model from Kelley and Tallman (2022) and the UTK-specific model were consistent with each other, with the exception of the nuchal crest inclusion in the UTK specific model. In the comparison of the metric models, the UTK specific model included the frontal chord, while the Kelley and Tallman (2022) metric model included biauricular breadth, nasal height, orbital height, bigonial breadth, and mandibular length. For the combined models, the UTK specific model included the nuchal crest, while the Kelley and Tallman (2022) combined model included the mental eminence, glabella occipital length, and mandibular length. The nuchal crest performed better with the UTK data, while the mandibular length performed better with the Kelley and Tallman (2022) data. This could be ascribed to the population variation between the two samples.

The somewhat reduced performance of Kelley and Tallman's (2022) models can be attributed to the heterogeneity of Kelley and Tallman's (2022) sample, which differs significantly from the more homogenous (i.e., majority European American) UTK sample. However, the results indicate that while sample-tailored models may perform better, population-inclusive models can achieve 80%or accuracy or above for some groups and be used in cases where ancestry is truly unknown. The lower performance for classifying AFAB individuals may be because the model has an easier time classifying males accurately because more robust features tend to show more distinct and consistent patterns. In the future, the data from this study will be combined with Kelley and Tallman's (2022) data to produce more robust population-inclusive models. Going forward, more work should be conducted to develop more robust population inclusive ASAB estimation models.

Contact

Victoria Conquest
Boston University Chobanian & Avedisian School of Medicine
Email: vconques@bu.edu

References