Sex estimation from virtual models: exploring the potential of stereolithic (STL) 3D crania models for morphoscopic trait scoring

Madeline Robles*† [a,b,d], Rachael M. Carew†1 [a,b], Carolyn Rando[c], Sherry Nakhaeizadeh [a,b] , Ruth M. Morgan [a,b]

[a] UCL Department of Security and Crime Science, 35 Tavistock Square, London, UK

[b] UCL Centre for the Forensic Sciences, 35 Tavistock Square, London, UK

[c] UCL Institute of Archaeology, 31-34 Gordon Square, London, UK

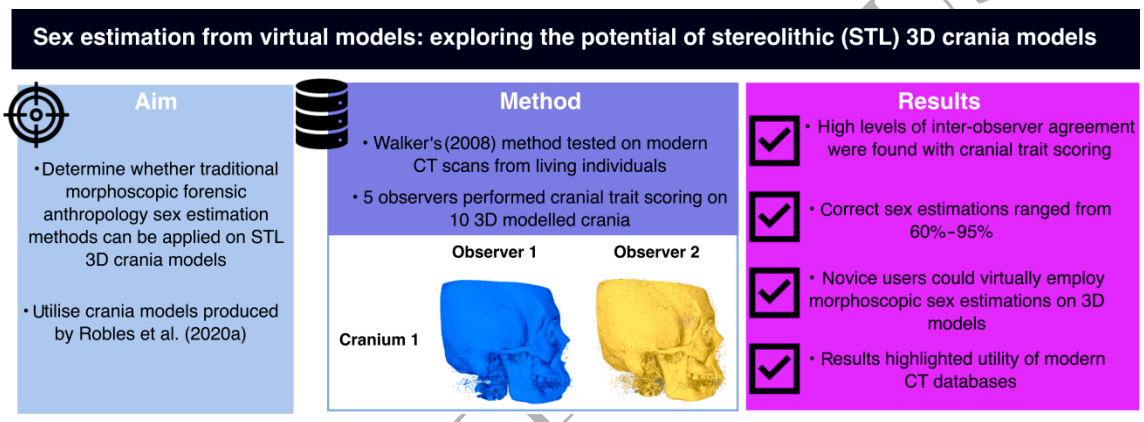
[d] School of Applied Sciences, College of Health, Science and Society, University of the West of England, Coldharbour Lane, Bristol, UK

madeline.robles@uwe.ac.uk *corresponding author
rachael.carew@coventry.ac.uk
c.rando@ucl.ac.uk
sherry.nakhaeizadeh@ucl.ac.uk
ruth.morgan@ucl.ac.uk

†Authors contributed equally

[1] Current Affiliation: School of Life Sciences, Faculty of Health and Life Sciences, Coventry University, Coventry, UK

Graphic abstract



Abstract

Modern computed tomography (CT) databases are becoming an accepted resource for the practice and development of identification methods in forensic anthropology. However, the utility of 3D models created using free and open-source visualisation software such as 3D Slicer have not yet been thoroughly assessed for morphoscopic biological profiling methods where virtual methods of analysis are becoming more common. This paper presents a study that builds on the initial findings from Robles et al. (2020) to determine the feasibility of estimating sex on STL 3D cranial models produced from CT scans from a modern, living UK population (*n*=80) using equation 2 from the Walker's (2008) morphoscopic method. Kendall's coefficients of concordance (KCC) indicated substantial agreement using cranial features scores in an inter-observer test and a video-inter-observer test. Fleiss's Kappa scores showed moderate agreement (0.50) overall between inter-observer sex estimations, and for observer sex estimations in comparison to recorded sexes (0.56). It was found that novice users could virtually employ morphoscopic sex estimation methods effectively on STL 3D cranial models from modern individuals. This study also highlights the potential that digital databases of modern living populations can offer forensic anthropology.

Key Points

- First example of Walker's (2008) method applied to a living UK population.
- Open-source software is a valuable resource for crime reconstruction approaches.
- Male scoring bias was observed in method application.
- Forensic anthropologists would benefit from virtual anthropology training to use and interpret 3D models.
- Digital databases offer more ethical, diverse, modern populations for future research.

Introduction

Historical skeletal collections and cemetery assemblages often act as a primary resource for forensic anthropologists in developing or testing biological profiling methods [1]. However, there are several drawbacks with relying solely on these collections. For example, these collections are not necessarily representative of contemporary (or indeed past) populations [1, 2], access to collections is extremely limited [3], and some raise ethical issues as a result of colonial antecedents and historical discriminatory practices [4]. The lack of appropriate, ethical and accessible collections consequently hinders the ability to test current methods used in forensic anthropology across forensically relevant modern global populations [3]. In recent years however, an alternative source for modern population data (derived from medical imaging databases) has been translated from its original medical purposes [5] for utilisation in forensic anthropology.

There has been growth in the exploration of the use of three-dimensional (3D) modelled bones from computed tomography (CT) data [6-10], and the use of medical imaging and virtual anthropology has been recognised as a suitable approach for developing and testing metric methods in forensic anthropology for direct applications to modern day populations [6, 7, 11-13]. However, there is little research that addresses the application and feasibility of forensic anthropological morphoscopic methods on 3D models of bones, which are arguably the most frequently used methods for sex and age estimations due to their ease of applicability [14, 15]. This study therefore further develops the work of Robles et al. [16] to determine the feasibility of estimating sex from virtual 3D cranial models (n=80) using the macromorphoscopic (hereafter morphoscopic) trait scoring method presented by Walker [17] using eight observers with various degrees of experience in employing forensic anthropological methods and 3D modelling.

Literature review

The use of modern imaging technologies to develop new approaches and methods within forensic anthropology applications is growing [7, 13]. Virtual 3D modelling of human anatomical structures has been established in forensic anthropology and is a tool that continues to be increasingly utilised [2, 6-10]. Although the accuracy of CT bone models has been confirmed in multiple studies [6, 18-20], virtual 3D modelling as

a visualisation approach is still in development [21]. Indeed, a large proportion of CT visualisation platforms have been applied and tested within forensic anthropology, including commercial platforms such as Mimics, Amira or Osirix [8], as well as free and open-source software, such as 3D Slicer or ITK-SNAP [22, 23]. In addition, the increased use of online platforms for training, teaching, and research is creating a new demand for the production and use of 3D models as primary teaching materials for anatomy, and forensic or biological anthropology applications [24, 25]. The increased use of online platforms and alternative teaching and research materials highlights the need for assessment of 3D models and their representation of anatomical structures.

However, the costs of licencing fees for commercial visualisation programmes (such as Mimics and Amira), and additional maintenance fees [8] can prove prohibitive for funding bodies and public sector organisations which reduces the accessibility of these tools [21]. A study by Abdullah et al. [22] identified no significant measurable differences in the 3D models produced between commercial and non-commercial visualisation platforms. However, to reliably implement free and open-source visualisation platforms such as 3D Slicer, there is a need to fully assess their capabilities within forensic anthropology, where virtual methods of analysis are likely to become even more essential [16]. Bertoglio et al. [26] investigated cranial CT models for morphoscopic analysis and found that the models were good representations overall, but also identified limitations such as areas of missing bone, missing anatomic details, and misinterpretation of bone anomalies as pathological lesions [26]. However, in their study the models from CT scans of dry bones were made, but only the volume renderings were then examined rather than a surface reconstruction made using segmentation [26]. Volume renders and surface reconstructions (including stereolithic (STL) models) are entirely different formats of 3D "models" that should always be explicitly identified to avoid misrepresentation. In terms of the issues identified related to missing bone or misinterpretation of anatomic details, Bertoglio et al. [26] suggested this could be resolved as software advances, however, such imaging anomalies will always be possible. Moreover, there is clearly a need for transparency in what specific models can achieve, and a place for training in medical imaging and 3D model reconstruction [27, 28].

The variability between populations has pushed studies to test the Walker method [17] across different populations [14, 29-31]. However, there are no published studies testing the Walker method [17] on a living UK population. Additionally, a number of studies have demonstrated the utility of CT data and/or 3D modelling and its pertinence in assessing morphoscopic differences to assist with sex estimation, such as for use with the maxillary sinus [32], the foramen magnum [33], or the pelvis [6, 34]. Considering cranial models specifically, 2D views and 3D volume reconstructions of the skull have been evaluated using general skull morphology [35] and using craniometrics [36]. Additionally, morphoscopic data have been obtained from volume renders [37]. However, the ability of anthropologists to utilise STL 3D models for traditional morphoscopic approaches (such as the Walker method [17]) and from UK modern population data is unknown. A step-by-step method for creating 3D models intended for those with minimal previous

experience [16] has demonstrated the potential accuracy of models created from CT data by a range of users with a reproducibility within 1–2 mm. As a next step, these models need to be further tested to establish whether they can be reliably used in forensic anthropology applications. Therefore, the study presented here sought to determine whether it was possible to apply traditional morphoscopic forensic anthropology sex estimation methods on the STL 3D cranial models produced by Robles et al. [16].

Material and methods

Participants

In Robles et al. [16] STL cranial models were produced from 20 clinical sinus CT scans (10 male and 10 female) by five observers. The crania were from living individuals of mean age 54.5 years (male 26–91 years, female 29–64 years). The cranial models were reconstructed using 3D Slicer version 4.9.0 (Brigham Women's Hospital, Boston, MA, USA) [38] following the method and scanning parameters outlined in Robles et al. [16]. Observers 1 and 2 had around 3 years of experience in 3D modelling and observers 3 and 4 had little to no prior experience. Observers 1–4 were all trained in forensic anthropology to master's degree level or higher. However, Observer 5 was not familiar with applying forensic anthropology methods and was thus excluded from this study. Figure 1 illustrates two of the crania (cranium 1 and 10) modelled by each of the original observers. All participants provided written informed consent prior to data collection.

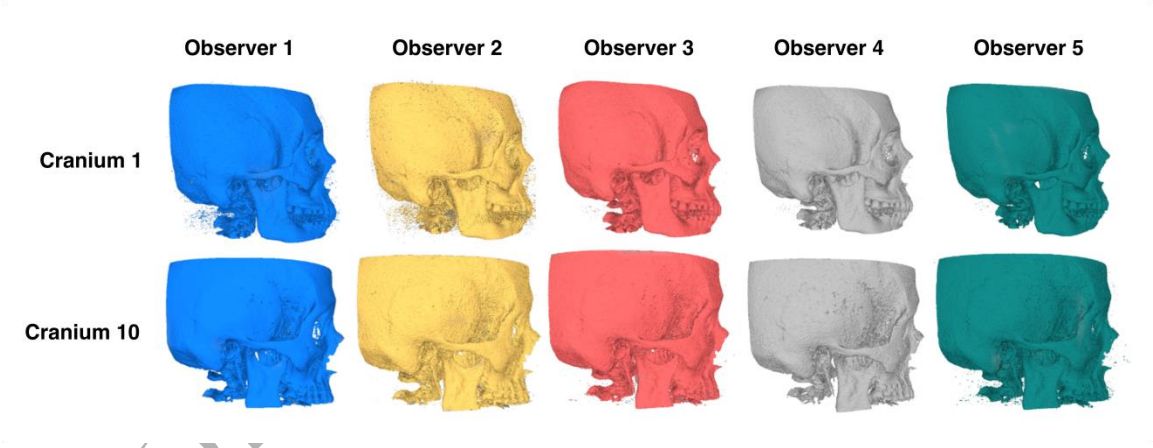


Figure 1. Matrix of two cranial stereolithic (STL) models (right lateral view) from all five observers in 3D Slicer.

Scores and equation

In this study, Observers 1–4 were asked to perform cranial trait scoring on the 3D models that they had created. Immediately after modelling, each observer re-loaded their STL cranial models into 3D Slicer using the "Data" module and scored the cranial morphoscopic traits using sex estimation methods based on Buikstra and Ubelaker (1994) taken from Walker [17]. The CT scans were obtained for viewing sinuses and as such did not include the complete crania, and only three cranial traits were consistently observable and scored—the mastoid process, supra-orbital margin, and glabella. The Walker method [17] allows for fragmentary or

incomplete skeletal elements to be used for sex estimations, as complete skeletal remains cannot be expected in forensic anthropological case work [17, 39].

Standard cranial trait scores of an ordinal scale of 1–5, as outlined by Walker [17] were used, with 1 typically representing more gracile ("female") features, and 5 more robust ("male") features. This method by Walker [17] was tested using American and British samples from the Hamann-Todd, Terry, and Saint Bride's Church collections and is regularly used across various populations [14]. Sex estimations for each crania were calculated using the cranial trait scores recorded by Observers 1–4 using logistic discriminant analysis equation number 2 from Walker [17] (Y=glabella×(-1.568)+mastoid×(-1.459)+7.434), which eliminates some of the subjectivity around the scoring. The cut-off value to discriminate between a male and female sex estimation is a score of zero using the equation [14]. Equation 2 uses the glabella and mastoid and was the only equation suitable for use with the traits available for this study. The sex estimations derived from the cranial trait scores were compared against the known recorded sex of each crania, with a percentage score for the number of correct classifications recorded for each observer.

Video observer test

Four additional observers (video Observers V1–V4) were recruited to further assess the robustness of the models through a "video observer test". The models were recorded using the screen recording function in QuickTime player ™ (.mov), where each cranium completed a 360° rotation about the lateral axis to provide full view of the cranial trait features in 3D Slicer. The full screen recording video was then shared with the video observers using a private link for the online platform YouTube. The incorporation of this "video observer test" created easy and remote access to the virtual models for four video observers to achieve a total of eight observers for this study.

The four video observers were forensic anthropology master's degree students (two currently studying and two graduates) who each had no previous experience of using virtual 3D models. These video observers scored each of the 20 cranial models produced by Observer 1, using the same scoring and sex estimation method as outlined above. The models produced by Observer 1 were considered the "gold standard" for comparisons and all of the models produced were confirmed as metrically accurate to each other and verified for use [16] .

Statistical analysis

The level of inter-observer agreement was evaluated. Data were analysed using Minitab® version 17.1 for Windows and prepared using Microsoft Excel version 16.23 for Mac (Microsoft, Redmond, WA, USA). Fleiss's Kappa [40] and Kendall's Coefficient of Concordance (KCC) were employed, with the strength

criteria from Landis and Koch [41] as a scale to assign agreement (as similarly used by Lewis and Garvin [14]) <0 = "poor", 0.0-0.20 = "slight", 0.21-0.40 = "fair", 0.41-0.60 = "moderate", 0.61-0.80 = "substantial", and 0.81-1.00 = "almost perfect agreement", meaning observer agreement is significantly greater than would be expected by chance. Kappa is appropriate for this dataset as it measures the degree of agreement for ordinal data (i.e., the cranial feature scores). Kappa is suitable for cases where multiple observers have assessed the same samples, and Fleiss's Kappa (rather than Cohen's Kappa) is used for more than two observers [42]. Additionally, Kendall's coefficients take ordering into consideration which results in not all misclassifications being treated equally [42]. For example, Kendall's coefficients consider that a score of 1 and 4 would have a higher degree of disagreement, than a score of 1 and 2. This ordering is appropriate for the cranial score data which are scored on a scale of 1–5.

Results

Cranial feature scores

The cranial feature scores from Observers 1–4 are presented in Table 1. In one case (Cranial 4) Observer 3 only scored the glabella. The results of the Kendall's coefficients of concordance (KCC) for each cranial feature were 0.68 for mastoid, 0.78 for supra-orbital margin, and 0.81 for the glabella, which indicated "substantial" to "almost perfect" agreement between the observers across the features using the Landis and Koch [41] classifications.

Table 1 Individual crania feature scores for Observers 1–4 (- denotes missing data).

Table 1 Individual cranial feature scores for Observers 1–4.

Crania	Mastoid process				Supra-orbital margin				Glabella			
	1	2	3	4	1	2	3	4	1	2	3	4
1	1	3	4	2	1	4	3	2	1	3	3	2
2	1	5	4	4	3	5	5	4	1	5	4	3
3	1	5	5	2	2	5	4	3	2	5	3	3
4	1	5	–	5	1	5	–	2	1	1	1	2
5	3	5	5	5	3	5	5	5	5	5	5	4
6	2	5	5	5	2	4	3	4	3	5	4	4
7	1	2	3	1	2	3	3	3	2	2	4	3
8	1	1	2	3	1	1	4	3	1	1	1	2
9	2	5	3	3	4	5	4	5	5	5	5	4
10	1	3	2	3	2	1	2	2	3	2	1	2
11	3	5	4	5	4	5	4	4	4	4	4	4
12	4	5	5	5	5	5	5	5	5	5	5	5
13	3	5	4	4	4	5	4	5	5	5	5	5
14	2	3	3	3	3	4	3	3	4	5	4	3
15	1	1	1	1	2	3	1	2	3	1	1	3
16	2	4	4	4	4	5	5	4	5	5	5	5
17	1	3	2	2	2	2	1	2	3	1	1	2
18	2	3	3	4	3	1	2	3	3	2	1	3
19	2	2	3	3	1	1	2	1	3	1	2	2
20	2	4	4	5	3	5	4	4	5	5	4	4

–: missing data.

Frequency plots of the cranial traits scores (Figure 2) illustrate the distribution of the score data. Observer 1 assigned the mastoid process with low scores more frequently than the higher scores, and Observer 2 assigned scores of 5 more often than other scores across traits. Generally, the scores have varied distribution.

Figure 2. Bar charts illustrating the frequency of the cranial trait scores (1–5) per trait and per observer.

Sex estimations

The cranial trait scores (Table 1) were used to obtain sex estimations using the Walker method [2] with equation 2 (Table 2).

Table 2. Sex estimation results for Observers 1–4 calculated from Walker (2008) Equation 2 [17] (prob m/f = probability male/female) and recorded sex for each crania 1–20.

Crania	Observer 1			Observer 2			Observer 3			Observer 4			Recorded
	Sex	Prob m (%)	Prob f (%)	Sex	Prob m (%)	Prob f (%)	Sex	Prob m (%)	Prob f (%)	Sex	Prob m (%)	Prob f (%)	
1	Female	1	99	Male*	84	16	Male*	96	4	Female	20	80	Female
2	Female	1	99	Male*	100	0	Male*	99	1	Male*	96	4	Female
3	Female	6	94	Male*	100	0	Male*	99	1	Male*	55	45	Female
4	Female	1	99	Male*	81	19	-*	-	-	Male*	95	5	Female
5	Male	99	1	Male	100	0	Male	100	0	Male	100	0	Male
6	Male	55	45	Male	100	0	Male	100	0	Male	100	0	Male
7	Female	6	94	Female	20	80	Male*	96	4	Female	22	78	Female
8	Female	1	99	Female	1	99	Female	5	95	Male*	52	48	Female
9	Male	97	3	Male	100	0	Male	99	1	Male	96	4	Male
10	Female	22	78	Male*	52	48	Female	5	95	Male*	96	4	Female
11	Male	96	4	Male	100	0	Male	99	1	Male	100	0	Male

12	Male	100	0	Male	100	0	Male	100	0	Male	100	0	Male
13	Male	99	1	Male	100	0	Male	100	0	Male	100	0	Male
14	Male	85	15	Male	99	1	Male	96	4	Male	84	16	Male
15	Female	22	78	Female	1	99	Female	1	99	Female	22	78	Female
16	Male	97	3	Male	100	0	Male	100	0	Male	100	0	Male
17	Female	22	78	Female	18	82	Female	5	95	Female	20	80	Female
18	Male	55	45	Male	52	48	Female*	18	82	Male	96	4	Male
19	Male*	55	45	Female	5	95	Male*	52	48	Male*	52	48	Female
20	Male	97	3	Male	100	0	Male	99	1	Male	100	0	Male
% correct	95			75			65			70			

*: incorrect sex assessment.

Accurate sex estimations were obtained in 65% to 95% of cases overall. Male crania were 90%–100% correctly estimated (average 95%), and female crania 40%–100% correct (average 58%).

Video observer scores

Four video observers scored 20 crania each in the video test. Individual cranial scores from the video observers are presented in Table 3. Kendall's coefficients of concordance for each cranial feature were 0.79 for mastoid, 0.76 for supra-orbital margin, and 0.84 for the glabella, which indicated "substantial" to "almost perfect" agreement between the video observers.

Table 3. Individual cranial feature scores for video Observers V1–V4.

Crania	Mastoid process				Supra-orbital margin				Glabella			
	V1	V2	V3	V4	V1	V2	V3	V4	V1	V2	V3	V4
1	2	4	2	2	2	4	2	3	3	3	3	2
2	4	5	3	2	4	4	2	2	4	4	2	2
3	3	5	2	3	4	3	3	4	5	3	1	2
4	5	5	2	2	3	2	1	3	2	1	1	1
5	5	5	5	4	5	5	5	5	5	5	5	5
6	5	5	4	4	4	4	2	3	5	4	3	3
7	1	2	1	1	3	2	1	4	3	3	2	3
8	3	2	1	1	2	2	3	2	1	1	1	1
9	3	4	4	2	5	5	5	5	5	5	5	5
10	4	4	3	3	3	2	1	3	3	3	2	2
11	5	5	5	4	4	4	3	2	4	4	2	2
12	5	5	5	4	5	5	3	4	5	5	5	5
13	4	5	4	4	4	5	5	5	5	5	5	5
14	3	4	2	3	4	4	3	3	4	4	2	1
15	1	3	1	2	3	2	1	4	3	3	2	2
16	4	5	4	3	5	5	4	5	5	4	5	4
17	2	3	3	2	3	2	2	3	3	2	1	3
18	3	4	2	3	3	3	3	4	2	2	1	1
19	1	4	3	3	1	1	1	2	1	1	2	2
20	5	5	4	4	4	4	4	4	5	5	4	3

Frequency plots of the cranial traits scores from the video observers (Figure 3) illustrate the distribution of the score data. The video observers appear to be assigning high scores more often than the lower scores.

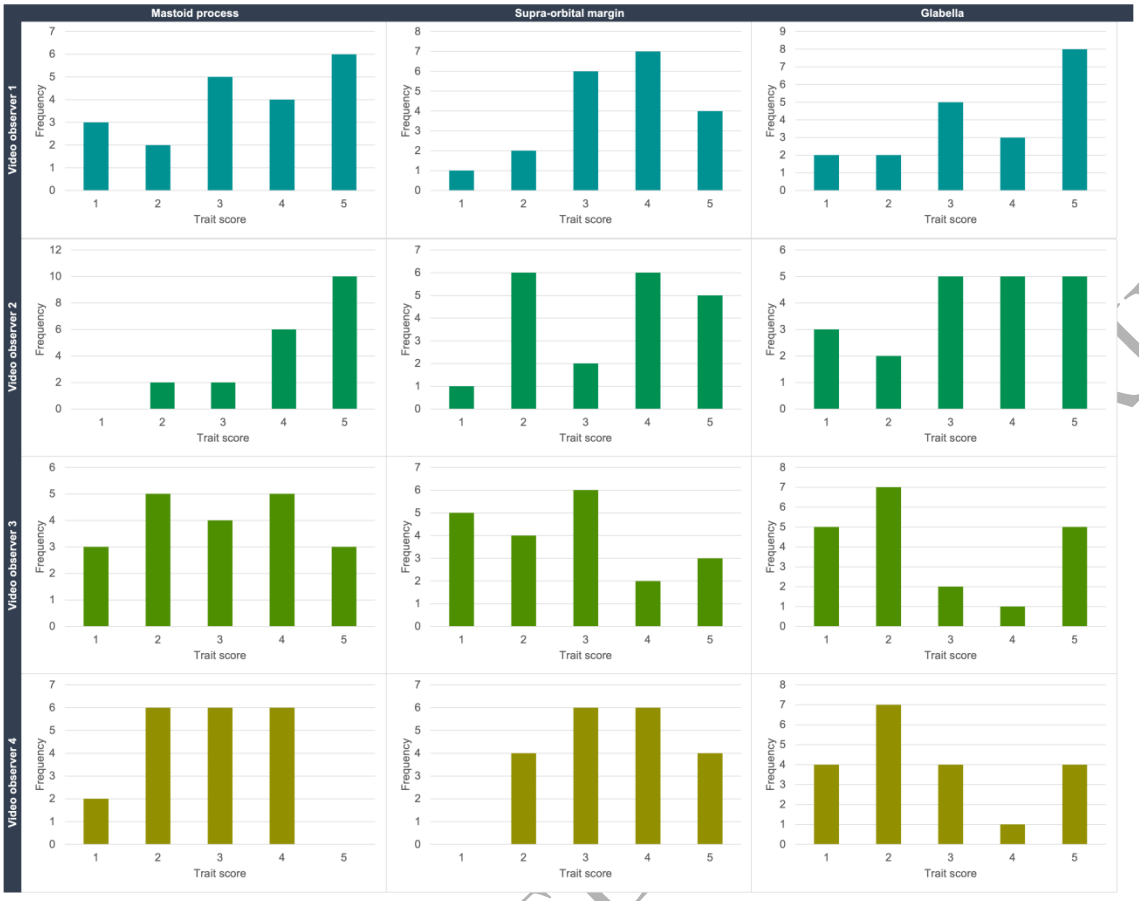


Figure 3. Bar charts illustrating the frequency of the cranial trait scores (1–5) per trait and per video observer.

Video observer sex estimations

The video observer cranial trait scores (Table 3) were used to obtain sex estimations using the Walker [17] method with Equation 2; the results are presented for video Observers V1–V4 in Table 4.

Table 4. Sex estimation results for video Observers V1–V4 calculated from Walker Equation 2 [17] (prob m/f = probability male/female) and recorded sex for each crania 1–20.

Crania	V1			V2			V3			V4			Recorded
	Sex	Prob m (%)	Prob F (%)	Sex	Prob m (%)	Prob F (%)	Sex	Prob m (%)	Prob F (%)	Sex	Prob m (%)	Prob F (%)	
1	Male*	55	45	Male*	96	4	Male*	55	45	Female	20	80	Female
2	Male*	99	1	Male*	100	0	Male*	52	48	Female	20	80	Female
3	Male*	99	1	Male*	99	1	Female	5	95	Male*	52	48	Female
4	Male*	95	5	Male*	81	19	Female	5	95	Female	5	95	Female
5	Male	100	0	Male	100	0	Male	100	0	Male	100	0	Male
6	Male	100	0	Male	100	0	Male	96	4	Male	96	4	Male

7	Female	22	78	Male*	55	45	Female	6	94	Female	22	78	Female
8	Female	18	82	Female	5	95	Female	1	99	Female	1	99	Female
9	Male	99	1	Male	100	0	Male	100	0	Male	100	0	Male
10	Male*	96	4	Male	96	4	Male*	52	48	Male*	52	48	Female
11	Male	100	0	Male	100	0	Male	95	5	Male	82	18	Male
12	Male	100	0	Male	100	0	Male	100	0	Male	100	0	Male
13	Male	100	0	Male	100	0	Male	100	0	Male	100	0	Male
14	Male	96	4	Male	99	1	Female	20	80	Female*	18	82	Male
15	Female	22	78	Male*	84	16	Female*	6	9	Female	20	80	Female
16	Male	100	0	Male	100	0	Male	100	0	Male	96	4	Male
17	Male*	55	45	Male*	52	48	Female	18	82	Male*	55	45	Female
18	Male	52	48	Male	82	18	Female*	5	95	Female*	18	82	Male
19	Female	1	99	Female	49	51	Male*	52	48	Male*	52	48	Female
20	Male	100	0	Male	100	0	Male	99	1	Male	96	4	Male
% correct	70			65			70			70			

*: incorrect sex assessment.

The cranial trait scores from the video observers correctly estimated the sex of the individual in 65%–70% of overall cases. Sex estimations were correctly classified in 80%–100% of cases for males (average 90%), and in 20%–45% of cases for females (average 40%).

Discussion

This study assessed the potential for using morphoscopic methods on 3D STL cranial models in forensic anthropology. Twenty different cranial models were examined by four observers who each performed cranial trait scoring following the morphoscopic method from Walker [17]. Four video observers also performed cranial trait scoring, but on videos of the 20 models produced by Observer 1.

A high level of agreement between morphoscopic feature scores was identified, indicating good agreement between the original observers, and between the video observers scores (KCC 0.68–0.84). Despite the high agreement, higher rates of incorrect feature scoring were observed at the start of modelling (e.g., Crania 1–4), which could potentially be explained by the observers familiarising themselves with this particular population, and its physical morphological traits. Moreover, some female crania do present with more robust traits (and *vice versa*), without knowing the variation present in the sample population, more "robust" female crania could be misinterpreted as possible male ones which is reflective of natural population variation [43]. Additionally, given the ordinal scoring system, trait scoring results alone cannot be interpreted for "accuracy".

In several instances, observers did not use a particular score at all with certain features, such as Observer 4 with the glabella (Figure 2), or video Observer 4 with the supra-orbital margin (Figure 3). Additionally, both

sets of observers frequently utilised the middle score of 3 (less so for Observer 2). These observations from the trait scoring could indicate uncertainty or a lack of confidence in utilising the method, or stem from a wider issue around lack of applicability of the method with the population used, and/or systematic bias towards certain scores. The possible influence of age on the cranial traits was not investigated in this study, but age has previously been discounted from playing a prominent role in cranial trait expression [43].

Published studies have examined the accuracy of traditional anthropological methods of establishing sex and found varying accuracy rates to be due to either populational differences, or simply to the experience of the observers [14] and their interpretations. The scores from the glabella exhibited higher agreement between observers, in concordance with previous studies that found particular features vary in their reliability [37, 44]. The level of agreement in this study was in line with published research reported by Langley et al. [44] using crania, by Villa et al. [10] for inter-observer agreement using pelvic features, and by Lesciotto and Doershuk [45] who found moderate to substantial inter-observer agreement (using pelvic features).

Additionally, the scores from the original observers (who were scoring their own models) resulted in accurate sex estimations for 65% to 95% of models using Equation 2 from Walker [17]. Similarly, the video observers (who were scoring the models from Observer 1), resulted in accurate sex estimations in 65%–70% of cases. These results are lower than can be seen in other studies, such as 91.8%–92.9% [37], 93.5% using dry skulls [44] and 82.9%–85.4% reported in Walker [17]. However, the equation used in this study only included two of the five possible scoring traits. Moreover, there appears to be some bias towards male scoring for both sets of observers, as male cranial scores resulted in correct sex classifications in 80%–100% of cases. A study by Oikonomopoulou et al. [29] reported similar accuracy differences between each sex, with males providing higher classification rates (above 90%) in contrast to the female sample (22.62%–61.36%). This could be explained by the observers having more familiarity with male skeletons, an issue stemming from assessing the robusticity and gracility of the 3D models, or potentially a wider methodological issue. It is salutary that there is evidence of male bias in forensic anthropology skeletal collections [39], in traditional method development [46], and even in modern machine learning approaches [47]. New population datasets and progressive approaches are needed to overcome such biases in forensic anthropology methodologies. Observer experience has previously been shown to influence the final sex classifications [14], however observer experience with sex estimation methods was not evaluated in this study as the aim was not to evaluate accuracy but feasibility. Higher rates of correct sex estimations were obtained from the original observer trait scores than the video observer scores, and this may be explained by the familiarity of the observers with virtual anthropology, indeed Observers 1 and 2 were familiar with 3D modelling or scoring 3D crania. Training in virtual anthropology and the development of new methods that are applicable to virtual anthropology approaches are vital.

The methods used were those typically taught in forensic anthropology programmes so that each observer was familiar with the procedures of the technique. However, the observers were not familiar with applying the methods to virtual 3D models or videos, which could have affected their ability to assess the cranial features. Three of the video observers remarked that the scoring process was difficult to implement visually without the use of touch, particularly for the supra-orbital margin, which may explain some of the variation seen in the scoring. Certainly, this reflects a limitation of virtual analysis, but also poses a wider question as to the transparency of decision making in evaluative interpretation [48-51] and specifically whether more tacit information elicited from "touch" can be incorporated into a framework for transparent evaluative decision making in a forensic science context [14, 52].

Overall, the models were successfully scored for cranial traits by all observers and the models, open-source software, and video productions provided straightforward, accessible platforms for conducting remote forensic anthropology analysis. The models used in this study were STL mesh models and not volume renderings, which is an important distinction that needs to be highlighted in research applications (see section "Literature review"). To comply with local ethical requirements, it was not possible to share the STL cranial files with participants. However, it was observed that the video test with a private link worked well as a user-friendly way to temporarily remotely share the models.

Scepticism about the utility of 3D modelling has focused on the misinterpretation of modelling artefacts as pathology or trauma [26]. Indeed, the models used in this study exhibited a degree of bone loss, for example, this can be seen around Pterion with Cranium 10 (Figure 1). Although there may be instances where the 3D model does not accurately represent minor morphological features, which could potentially result in erroneous trauma and pathological identifications [26, 53], this highlights the importance of training in 3D CT modelling for forensic anthropologists. Indeed, users should understand that any missing data may be the result of CT slicing or thresholding errors, and thereby avoid misinterpreting artefacts as pathology or trauma. Moreover, these findings emphasise the need for training and establishing quality control protocols in model development, and inter-observer testing for forensic reconstructions.

The opportunity to apply the capabilities of modern imaging technologies creates new avenues of research where visual procedures in the interpretation of skeletal remains could be further enhanced using methods that may offer a less time-consuming approach (for example over manual maceration techniques), and imaging approaches facilitate remote and immediate access to scan data or virtual models. Further, using virtual anthropology and modern scan data from a living population, supports a more ethical approach than traditional osteological approaches that can avoid maceration of human remains, overhandling of skeletal collections, and colonialism and historical discriminatory practices [4]. While there are associated benefits to using virtual anthropology, it is also vital to understand the underlying factors that play a role in the interpretation of current and new methods in virtual environments, including testing for the reliability and

accuracy of the applicability of 3D STL cranial models in a forensic context. However, alternative ethical issues have arisen and are starting to be explored concerning the production of 3D models [54, 55]. Given the existing restrictions that can make physical access to skeletal collections difficult, there is clearly huge potential for 3D models to increase accessibility to collections through digital databases and radiographic imaging. For example, CT scanning is routinely carried out prior to autopsy in several institutes, which increases the datasets of modern populations available that may be suitable for research purposes [56] in addition to clinical datasets of living patients. Virtual anthropology offers an alternative pathway for data collection within forensic anthropology when access to traditional skeletal collections is either limited, or not possible. Therefore, traditional methods for establishing a biological profile must be further tested on virtual models to determine feasibility, as well using contemporary population datasets with contemporary discriminate function equations to improve sex estimation classifications systems. This initial study has only begun to test the feasibility of STL 3D models and highlights the need for further research to be conducted in order to establish the scope of using traditional morphoscopic methods on different skeletal elements.

The main aim of this study was to determine if it was possible to visually assess STL 3D cranial models from a modern UK population, but not to assess the accuracy of the sex estimation results. Therefore, the sex estimation results found in this study were reasonable as an indicator of sex estimation accuracy for the purpose of assessing the usability of 3D crania. Good compatibility with the sex estimation scoring method adds further weight to the robustness of the cranial models produced previsouly [16]. The results from this study thus add weight to the suitability of the STL 3D cranial models produced by Robles et al. [16] for morphoscopic analysis.

Conclusions

This study has demonstrated that it was possible to apply a traditional morphoscopic forensic anthropology sex estimation method on the STL 3D cranial models produced by Robles et al. [16]. This study is the first (to our knowledge) to test the Walker method [17] on STL 3D models produced from CT data from a living UK population. Levels of inter-observer agreement were found with cranial trait scoring, and correct sex estimations ranged from 65%–95% for both sets of observers, albeit with probable bias towards male scoring. High percentages of correct male classification were observed, with lower female classification rates.

Complementary studies are needed to assess traditional macromorphoscopic methods on other skeletal STL models such as the pubic symphysis and auricular surface from a variety of modern populations. Potential male bias in anthropology teaching and/or skeletal collections could be overcome with the utilisation of modern 3D models. A comparison between interpretations made using volume renderings and those made using STL surface reconstructions would also be useful to assess whether there is any potential impact from these two digital approaches.

The ability to use free software such as 3D Slicer to view STL 3D models for morphoscopic trait scoring is important for forensic science applications in a field where funding is often very limited. It is also salient to consider how these tools will enable the development of digital databases that not only offer access to broader and more diverse populations for practitioners and researchers, but also opens up new areas of research that can be carried out with modern CT data where modern day populations are particularly relevant, as in forensic anthropology reconstructions.

Acknowledgements

We would like to extend our gratitude to each of the observers that volunteered their time to take part in this research and to the Picture Archiving and Communications Department at University College London Hospital (UCLH) for providing the CT scan data. We would also like to thank the anonymous reviewers for their helpful and insightful feedback.

Authors' contributions

Madeline Robles was a key contributor to the conceptualisation, method design, CT and 3D data collection, data analysis, project coordination and drafted manuscript. Rachael M. Carew was responsible for the conceptualisation, method design, data analysis and drafted manuscript. Carolyn Rando contributed to the conceptualisation, method design and project supervision. Sherry Nakhaeizadeh participated in editing the manuscript and project supervision. Ruth M. Morgan was a key contributor to the conceptualisation, method design and project supervision. All authors contributed to the final text and approved it.

Compliance with ethical standards

Approval was received from the Health Research Authority (HRA) and deemed exempt from requiring NHS REC approval by the HRA.

Disclosure statement

The authors report there are no competing interests to declare.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

1. Komar DA, Grivas C. Manufactured populations: what do contemporary reference skeletal collections represent? A comparative study using the Maxwell Museum documented collection. Am J Phys Anthropol. 2008;137:224-233.
2. Hisham S, Franklin D. Transposition of the Suchey–Brooks and spheno-occipital synchondrosis fusion methods onto computed tomographic images: review and future prospects. Forensic Imaging, 2020. 21.
3. Stock, M.K., et al., The importance of processing procedures and threshold values in CT scan segmentation of skeletal elements: An example using the immature os coxa. Forensic Sci Int. 2020. 309: p. 110232.
4. Biers, T., Rethinking Purpose, Protocol, and Popularity in Displaying the Dead in Museums, in Ethical Approaches to Human Remains, K. Squires, D. Errickson, and N. Márquez-Grant, Editors. 2020: Switzerland. p. 239-263.
5. D'Arcy, G., N. Marquez-Grant, and D.W. Lane, Baggage scanners and their use as an imaging resource in mass fatality incidents. Int J Legal Med. 2020;134:1419-1429.
6. Decker, S.J., et al. Virtual determination of sex: metric and nonmetric traits of the adult pelvis from 3D computed tomography models. J Forensic Sci. 2011;56:1107-1114.
7. Robles, M., C. Rando, and R.M. Morgan. The utility of three-dimensional models of paranasal sinuses to establish age, sex, and ancestry across three modern populations: A preliminary study. Australian Journal of Forensic Sciences. 2020:1-20.
8. Julieta G. García-Donas, Suna Ors, Ercan Inci, et al. Sex Estimation in a Turkish Population Using Purkait's Triangle: A Virtual Approach by 3-Dimensional Computed Tomography (3D-CT). Forensic Sci Res. 2022;7:97–105.
9. Franklin, D., L. Swift, and A. Flavel, "Virtual anthropology" and radiographic imaging in the Forensic Medical Sciences. Egypt J Forensic Sci. 2016;6:31-43.
10. Min Zhang. Forensic Imaging: A Powerful Tool in Modern Forensic Investigation. Forensic Sci Res. 2022;7:385–392.
11. Christensen, A., et al., The Use of X-ray Computed Tomography Technologies in Forensic Anthropology. Forensic Anthropology, 2018;1:124-140.
12. Brough, A., et al., The benefits of medical imaging and 3D modelling to the field of forensic anthropology positional statement of the members of the forensic anthropology working group of the International Society of Forensic Radiology and Imaging. J. Forensic Radiol. Imaging, 2019;18:18-19.
13. Bewes, J., et al., Artificial intelligence for sex determination of skeletal remains: Application of a deep learning artificial neural network to human skulls. J Forensic Leg Med. 2019;62:40-43.
14. Lewis, C.J. and H.M. Garvin, Reliability of the Walker Cranial Nonmetric Method and Implications for Sex Estimation. J Forensic Sci. 2016;61:743-751.

15. Klales, A. Current practices in forensic anthropology for sex estimation in unidentified, adult individuals. in Annual Meeting of the American Academy of Forensic Sciences. 2013. Washington, DC: Colorado Springs, CO: American Academy of Forensic Sciences.
16. Robles, M., et al., A step-by-step method for producing 3D crania models from CT data. J. Forensic Imaging, 2020. 23.
17. Walker, P.L., Sexing skulls using discriminant function analysis of visually assessed traits. Am J Phys Anthropol. 2008;136: 39-50.
18. Carew, R.M., et al., Accuracy of computed radiography in osteometry: A comparison of digital imaging techniques and the effect of magnification. J Forensic Radiol Imaging, 2019;19:100348.
19. Brough, A.L., et al., Anthropological measurement of the juvenile clavicle using multi-detector computed tomography--affirming reliability. J Forensic Sci. 2013;58:946-951.
20. Stull, K.E., et al., Accuracy and reliability of measurements obtained from computed tomography 3D volume rendered images. Forensic Sci Int. 2014;238:133-140.
21. Simmons-Ehrhardt, T., Open osteology: Medical imaging databases as skeletal collections. Forensic Imaging, 2021. 26.
22. Abdullah, J.Y., et al., Comparison of STL skull models produced using open-source software versus commercial software. Rapid Prototyping Journal. 2019;25:1585-1591.
23. Buser, T.J., et al., The Natural Historian's Guide to the CT Galaxy: Step-by-Step Instructions for Preparing and Analyzing Computed Tomographic (CT) Data Using Cross-Platform, Open Access Software. Integrative Organismal Biology, 2020. 2(1).
24. Sleiwah, A., et al., COVID-19 lockdown learning: The uprising of virtual teaching. Journal of Plastic, Reconstructive & Aesthetic Surgery. 2020;73:1575 - 1592.
25. Thompson, T.J.U., et al., Forensic undergraduate education during and after the COVID-19 imposed lockdown: Strategies and reflections from India and the UK. Forensic Sci Int, 2020. 316: p. 110500.
26. Bertoglio, B., et al., Pitfalls of Computed Tomography 3D Reconstruction Models in Cranial Nonmetric Analysis. J Forensic Sci. 2020.
27. Marquez-Grant, N., An overview of age estimation in forensic anthropology: perspectives and practical considerations. Ann Hum Biol. 2015;42:308-322.
28. Carew, R.M. and D. Errickson, Imaging in Forensic Science: Five Years On. J. Forensic Radiol Imaging. 2019;16:24-33.
29. Oikonomopoulou, E.K., E. Valakos, and E. Nikita, Population-specificity of sexual dimorphism in cranial and pelvic traits: evaluation of existing and proposal of new functions for sex assessment in a Greek assemblage. Int J Legal Med. 2017;131:1731-1738.
30. Soficaru, A., et al., Evaluation of discriminant functions for sexing skulls from visually assessed traits applied in the Rainer Osteological Collection (Bucharest, Romania). Homo. 2014;65:464-475.
31. Jilala, W., et al., Sexing contemporary Tanzanian skeletonized remains using skull morphology: A test of the walker sex assessment method. Forensic Science International: Reports, 2021. 3.

32. Abasi, P., et al., Comparison of accuracy of the maxillary sinus area and dimensions for sex estimation lateral cephalograms of Iranian samples. J. Forensic Radiol. Imaging, 2019;17:18-22.
33. Lashin, H.I., B.S. Eldeeb, and M.M. Ghonem, Sex identification from foramen magnum using computed tomography scanning in a sample of Egyptian population. J. Forensic Radiol Imaging. 2019;19:100341.
34. Franchi, A., et al., The prospects for application of computational anatomy in forensic anthropology for sex determination. Forensic Sci Int. 2019;297:156-160.
35. Grabherr, S., et al., Estimation of sex and age of "virtual skeletons"--a feasibility study. Eur Radiol. 2009;19:419-429.
36. Simmons-Ehrhardt, T.L., C.J. Ehrhardt, and K.L. Monson, Evaluation of the suitability of cranial measurements obtained from surface-rendered CT scans of living people for estimating sex and ancestry. J. Forensic Radiol. Imaging, 2019;19:100338.
37. Dereli, A.K., et al., Sex determination with morphological characteristics of the skull by using 3D modeling techniques in computerized tomography. Forensic Sci Med Pathol. 2018;14:450-459.
38. Fedorov, A., et al., 3D Slicer as an image computing platform for the Quantitative Imaging Network. Magn Reson Imaging. 2012;30:1323-1341.
39. Spradley, M.K. and R.L. Jantz, Sex estimation in forensic anthropology: skull versus postcranial elements. J Forensic Sci. 2011;56:289-296.
40. Fleiss, J.L., Measuring nominal scale agreement among many raters. Psychological Bulletin. 1971;76:378-382.
41. Landis, J.R. and G.G. Koch, The Measurement of Observer Agreement for Categorical Data. Biometrics. 1977;33:159-174.
42. Minitab. Kappa statistics and Kendall's coefficients. 2019 2019 15th January 2021]; Available from: https://support.minitab.com/en-us/minitab/18/help-and-how-to/quality-and-process-improvement/measurement-system-analysis/supporting-topics/attribute-agreement-analysis/kappa-statistics-and-kendall-s-coefficients/.
43. Garvin, H.M., S.B. Sholts, and L.A. Mosca, Sexual dimorphism in human cranial trait scores: effects of population, age, and body size. Am J Phys Anthropol. 2014;154:259-269.
44. Langley, N.R., B. Dudzik, and A. Cloutier, A Decision Tree for Nonmetric Sex Assessment from the Skull. J Forensic Sci. 2018;63:31-37.
45. Lesciotto, K.M. and L.J. Doershuk, Accuracy and Reliability of the Klales et al. (2012) Morphoscopic Pelvic Sexing Method. J Forensic Sci. 2018;63:214-220.
46. Spradley, M.K., et al., Demographic change and forensic identification: problems in metric identification of Hispanic skeletons. J Forensic Sci. 2008;53:21-28.
47. Nikita, E. and P. Nikitas, On the use of machine learning algorithms in forensic anthropology. Leg Med (Tokyo). 2020;47: 101771.

48. Morgan, R.M., Conceptualising forensic science and forensic reconstruction. Part II: The critical interaction between research, policy/law and practice. Sci Justice, 2017. 57(6): p. 460-467.
49. Morgan, R.M., Conceptualising forensic science and forensic reconstruction. Part I: A conceptual model. Sci Justice. 2017;57:455-459.
50. Earwaker, H., et al., A cultural change to enable improved decision-making in forensic science: A six phased approach. Sci Justice. 2020;60:9-19.
51. Georgiou, N., R.M. Morgan, and J.C. French, Conceptualising, evaluating and communicating uncertainty in forensic science: Identifying commonly used tools through an interdisciplinary configurative review. Sci Justice. 2020;60:313-336.
52. Nakhaeizadeh, S., I.E. Dror, and R.M. Morgan, Cognitive bias in forensic anthropology: Visual assessment of skeletal remains is susceptible to confirmation bias. Sci Justice. 2014;54:208-214.
53. Colman, K.L., et al., Virtual forensic anthropology: The accuracy of osteometric analysis of 3D bone models derived from clinical computed tomography (CT) scans. Forensic Sci Int. 2019;304:109963.
54. Carew, R.M., et al., Exploring public perceptions of creating and using 3D printed human remains. Forensic Science International: Reports. 2023.
55. Carew, R.M., J. French, and M.R. M., An ethical framework for the creation and use of 3D printed human remains in crime reconstruction. Forensic Sci Int: Reports. 2023.
56. Villa, C., J. Buckberry, and N. Lynnerup, Evaluating osteological ageing from digital data. J Anat. 2019;235:386-395.