Virtual forensic anthropology: The accuracy of osteometric analysis of 3D bone models derived from clinical computed tomography (CT) scans

Kerri L. Colman[a],*, Hans H. de Boer[b,c], Johannes G.G. Dobbe[d], Niels P.T.J. Liberton[e], Kyra E. Stull[f,g], Maureen van Eijnatten[h,i], Geert J. Streekstra[d,j], Roelof-Jan Oostra[a], Rick R. van Rijn[c,j] and Alie E. van der Merwe[a]

[a] Amsterdam UMC, University of Amsterdam, Department of Medical Biology, section Clinical Anatomy and Embryology, Meibergdreef 9, Amsterdam, the Netherlands
[b] Amsterdam UMC, University of Amsterdam, Department of Pathology, Meibergdreef 9, Amsterdam, the Netherlands
[c] Netherlands Forensic Institute, Department of Forensic Medicine, Den Haag, the Netherlands
[d] Amsterdam UMC, University of Amsterdam, Department of Biomedical Engineering and Physics, Amsterdam Movement Sciences, Meibergdreef 9, Amsterdam, the Netherlands
[e] Amsterdam UMC, University of Amsterdam, Medical Technology, 3D Innovation Lab, de Boelelaan 1117, Amsterdam, the Netherlands
[f] University of Nevada, Reno, Department of Anthropology, Reno, NV, USA
[g] University of Pretoria, Faculty of Health Sciences, Department of Anatomy, Pretoria, South Africa
[h] Amsterdam UMC, University of Amsterdam, Department of Oral and Maxillofacial Surgery & 3D Innovation Lab, de Boelelaan 1117, Amsterdam, the Netherlands
[i] Centrum Wiskunde & Informatica (CWI), Computational Imaging group, Amsterdam, the Netherlands
[j] Amsterdam UMC, University of Amsterdam, Department of Radiology, Meibergdreef 9, Amsterdam, the Netherlands

*Correspondence to: k.l.colman@amsterdamumc.nl

E-mail addresses: k.l.colman@amsterdamumc.nl (K.L.Colman), h.h.deboer@amsterdamumc.nl (H.H.deBoer), j.g.g.dobbe@amsterdamumc.nl (J.G.G.Dobbe), n.liberton@amsterdamumc.nl (N.P.T.J.Liberton), kstull@unr.edu (K.E.Stull), m.vaneijnatten@amsterdamumc.nl (M.vanEijnatten), g.j.streekstra@amsterdamumc.nl (G.J.Streekstra), r.j.oostra@amsterdamumc.nl (R.-J.Oostra), r.r.vanrijn@amsterdamumc.nl (R.R.vanRijn), a.e.vandermerwe@amsterdamumc.nl (A.E.vanderMerwe).

Highlights

- Virtual bones generated from 'clinical' CT scans are larger in size than the dry skeletal element.
- Correctly locating landmarks on virtual bones differs per modality and landmark.
- Methods derived from virtual bones may not always be applied to dry bones.

Abstract

Clinical radiology is increasingly used as a source of data to test or develop forensic anthropological methods, especially in countries where contemporary skeletal collections are not available. Naturally, this requires analysis of the error that is a result of low accuracy of the modality (i.e. accuracy of the segmentation) and the error that arises due to difficulties in landmark recognition in virtual models. The cumulative effect of these errors ultimately determines whether virtual and dry bone measurements can be used interchangeably.

To test the interchangeability of virtual and dry bone measurements, 13 male and 14 female intact cadavers from the body donation program of the Amsterdam UMC were CT scanned using a standard patient scanning protocol and processed to obtain the dry os coxae. These

were again CT scanned using the same scanning protocol. All CT scans were segmented to create 3D virtual bone models of the os coxae (‘dry’ CT models and ‘clinical’ CT models). An Artec Spider 3D optical scanner was used to produce gold standard ‘optical 3D models’ of ten dry os coxae.

The deviation of the surfaces of the 3D virtual bone models compared to the gold standard was used to calculate the accuracy of the CT models, both for the overall os coxae and for selected landmarks. Landmark recognition was studied by comparing the TEM and %TEM of nine traditional inter-landmark distances (ILDs). The percentage difference for the various ILDs between modalities was used to gauge the practical implications of both errors combined.

Results showed that ‘dry’ CT models were 0.36–0.45 mm larger than the ‘optical 3D models’ (deviations −0.27 mm to 2.86 mm). ‘Clinical’ CT models were 0.64–0.88 mm larger than the ‘optical 3D models’ (deviations −4.99 mm to 5.00 mm). The accuracies of the ROIs were variable and larger for ‘clinical’ CT models than for ‘dry’ CT models. TEM and %TEM were generally in the acceptable ranges for all ILDs whilst no single modality was obviously more or less reliable than the others. For almost all ILDs, the average percentage difference between modalities was substantially larger than the average percentage difference between observers in ‘dry bone’ measurements only.

Our results show that the combined error of segmentation- and landmark recognition error can be substantial, which may preclude the usage of ‘clinical’ CT scans as an alternative source for forensic anthropological reference data.

Keywords

Three-dimensional imaging
Forensic anthropology
Accuracy
Reliability
Pelvis
Computed tomography

1. Introduction

In many countries, especially in Europe, contemporary skeletal collections do not exist and therefore development and further validation of forensic anthropological methods for contemporary populations is impossible. In search of a solution, several researchers have used radiological data, such as computed tomography (CT) scans, as an alternative source for skeletal data [[1], [2], [3], [4], [5], [6], [7], [8], [9]]. By generating 3D virtual bone models from clinical scan data, contemporary virtual skeletal population samples can be created for the development and/or testing of methods. However, before 3D virtual bone models derived from clinical CT scans can be considered a reliable alternative source of population data, more information is needed regarding the accuracy of these models. Specifically, the degree to which 3D virtual bone models derived from clinical CT scans agree in shape and size to their dry skeletal counterparts.

While some studies have explored the accuracy of 3D virtual bone models and produced favorable results, these virtual models were based on postmortem CT data [10] or CT scans of dry skeletal- [[11], [12], [13], [14], [15]] or partially decomposed remains [16]. Scans of such elements likely result in a better image quality because higher levels of radiation can be used, and there is a reduced amount of soft tissue to contend with [17]. The optimistic findings in these studies may be biased if using clinical CT scans. Therefore, the accuracy of 3D virtual bone models derived from clinical CT data is still largely unknown.

Moreover, the aforementioned accuracy studies all used standard linear measurements as a means to identify differences in dimensions (referred to as size differences) between the modalities (the actual dry skeletal elements vs. 3D virtual bone models). Although linear measurements may be a gauge for the accuracy of a 3D virtual bone model, these measurements incorporate both the observer error as well as the actual size differences. This means that the differences observed between measurements taken from dry skeletal elements and 3D virtual models may not only be the result of size differences/similarities between the modalities but may also be from difficulties associated with landmark recognition.

Optical scanning may offer an approach to remove observer error and/or landmark recognition, thereby ensuring differences between the models is purely due to imaging. Optical scanning has less than a 0.05 mm difference between the true and virtual objects, which is an accuracy that far exceeds the virtual reproductions that a CT scanner can achieve [18]. Therefore, by comparing the clinical CT derived 3D bone models to the 3D bone models generated by optically scanning the dry bone counterparts, the accuracy can be quantified without the influence of landmark recognition.

In this study we aim to evaluate the accuracy of 3D virtual models derived from clinical CT data using optical scans and to investigate the influence of landmark recognition from these models. The implication of the compounded effect of the two aforementioned sources of variation (virtual modeling by segmentation of the bone, and landmark recognition) will be investigated and discussed from a forensic anthropological point of view.

2. Materials and methods

The complexity of creating 3D virtual models from CT data varies per skeletal element. In order to ensure that the results obtained in this study can be representative of the most complex modelling processes, a decision was made to focus on the pelvis. Its low signal-to-noise ratio in CT scans adds complexity to the virtual modelling process and thus represents a ‘worst-case scenario’.

Twenty-seven (13 male, 14 female) randomly selected, fully intact cadavers obtained from the body donation program of the Amsterdam UMC, location Academic Medical Center (AMC), University of Amsterdam, Department of Medical Biology, section Clinical Anatomy and Embryology were included in the study. Donation and use of cadavers were done in accordance with Dutch legislation and the regulations of the medical ethical committee of the Amsterdam UMC, location AMC. All individuals were older than 50 years of age at the time of death with an age range of 52–94 years for males and 64–100 years for females.

Virtual bone models were created and compared to their dry bone counterparts, either by models created through optical scanning or by means of direct comparison through linear measurements to investigate: A) the accuracy of clinical CT derived 3D virtual bone models in comparison to their dry skeletal counterparts, B) the influence of the aforementioned models on landmark recognition, and C) the practical implication of both errors combined (virtual modeling by segmentation of the bone, and landmark recognition).

To obtain the dry skeletal counterparts and their 3D virtual bone models, the following procedures were followed. A CT scan was made of the pelvis of all cadavers while fully intact using a standard patient/clinical scanning protocol (120 kV, 150 mAs, slice thickness 0.9 mm, increment 0.45 mm, reconstruction kernel D) on a Philips Brilliance 64 scanner (Philips Medical Systems, Best, The Netherlands). Given the specific clinical scanning protocol, these scans are hereafter referred to as the 'clinical' CT scans. Thereafter, the pelves were macerated by removing the majority of the soft tissue and letting the remains simmer in warm water (80 °C) for approximately 12 h. After which the dry skeletal remains were left to air dry. Following maceration, a 'dry' CT scan was made of the resulting dry os coxae (further referred to as the 'dry bones') using the same scanning protocol as mentioned above.

Both sets of scan data (i.e. the 'dry' CT and 'clinical' CT) were segmented using dedicated in-house research software to create 3D virtual bone models. These two sets of 3D virtual bone models will further be referred to as the 'dry' CT models and 'clinical' CT models, respectively. More in depth information regarding the software package and the segmentation process can be found in Dobbe et al. (2011) [19], Dobbe et al. (2018) [20] and Colman et al. (2017) [21].

Five sets of dry os coxae bones (n = 10) were scanned using an Artec Spider 3D optical scanner (Artec 3D, Luxembourg) to investigate the accuracy of the CT based 3D virtual bone models. Given the optical scanner's documented minimal error (<0.05 mm) and high resolution (~0.1 mm), the 3D bone models (further referred to as the 'optical 3D' models) produced from optically scanning the dry bone elements are considered to be the gold standard [22]. The 'optical 3D' models can subsequently be used as a reference to effectively evaluate the accuracy of the 'dry' CT models and the 'clinical' CT models without compounding the results with human error via landmark recognition.

2.1. Part A: accuracy of 3D virtual modeling

The 'optical 3D' models were virtually superimposed on both of the 'dry' and 'clinical' CT models using GOM Inspect® software (GOM Inspect® v8, GOM mbH, Braunschweig, Germany) to study the accuracy of the 'dry' CT models and the 'clinical' CT models, without the influence of error due to landmark recognition. Superimposition was achieved by using a Gaussian best-fit approach, which is done by minimizing the sum squared deviations between selected points and the given surfaces of the polygon meshes.

Additionally, regions of interest (ROI) were selected at the sites of eight well described landmarks (LMs) (Table 1) [23,24]. These LMs were chosen based on two criteria: 1) they are commonly used in traditional metric analysis of the os coxa [23,25] and 2) they represent

areas with variant levels of 3D virtual bone model precision as previously described by Colman et al. (2017) [26]. The ROI were manually selected in the software program on each of the 'optical 3D' models using a sphere that measured 4 mm in diameter. The centroid of the mesh points enclosed by the spherical ROI was recorded and was used for comparison with the corresponding 'dry' and 'clinical' CT models.

Table 1. Description of the eight pre-selected landmarks (LM) from which nine inter-landmark distances (ILD) were measured (as shown in Fig. 1) [23]. In brackets: the category/type of landmark according to Bytheway et al. (2010) [24]. These are used to investigate the influence of CT derived virtual bone models on landmark recognition.

Number	Definitions of Landmarks (LM) [23,24]
LM 1	Apex of the posterior superior iliac spine *(Traditional/Type 2)*
LM 2	Apex of the posterior inferior iliac spine *(Traditional/Type 2)*
LM 3	Farthest point of ischial curve from the center of the obturator foramen *(Extremal-Fuzzy/Type 3)*
LM 4	Most superior point on the superior edge of the medial aspect of the pubic symphysis *(Traditional/Type 2)*
LM 5	Most posterior point of the obturator foramen *(Constructed)*
LM 6	Apex of the anterior inferior iliac spine *(Traditional/Type 2)*
LM 7	Apex of the anterior superior iliac spine *(Traditional/Type 2)*
LM 8	Most superior point of the iliac crest, in measuring position *(Traditional/Type 2)*

The agreement between each CT-derived 3D virtual bone model ('dry' CT- and 'clinical' CT) and its corresponding gold standard ('optical 3D' model) was represented by a distance map, showing the perpendicular distance from the CT-derived model to the gold standard model. The distribution of this parameter, the arithmetic mean (in mm) is calculated for each distance map in both a negative and positive direction. These were calculated for the overall os coxae (left and right combined, sacrum excluded) and for each manually selected region of interest (ROI).

2.2. Part B: landmark recognition

To assess the human error component associated with landmark recognition (i.e. intra- and inter-observer error), nine inter-landmark distances (ILDs) were measured (Fig. 1) using the aforementioned selected LMs (Table 1 and Fig. 1) [24]. All ILDs were measured on the left os coxa from the 'dry bones', as well as on the left element in the corresponding sets of 'dry' CT- and 'clinical' CT virtual bone models. 'Dry bone' measurements were performed using a digital sliding caliper. Measurements on the 3D virtual bone models were taken using an integrated measuring tool in the in-house research software [19]. All measurements were rounded off to one tenth of a millimeter. Each measurement was taken twice by two independent observers, both with equal experience in the field, with a minimum of one-week lapse between repeated measurements. Measurements could be obtained on all 27 left os coxae.

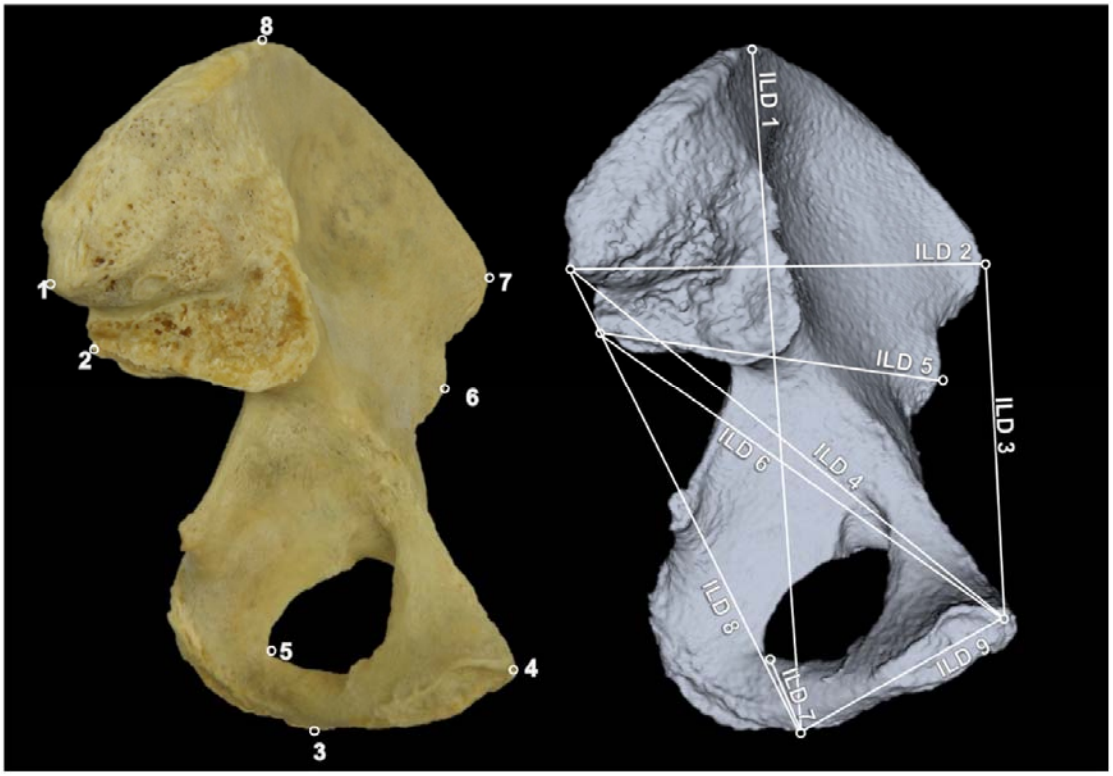

Fig. 1. Graphic illustration of the eight (numbered 1–8) pre-selected landmarks (LM), as defined in Table 1 of this paper and the associated inter-landmark distances (ILD) measured to investigate the influence of clinical CT derived virtual bone models on landmark recognition. ILD 1 = LM 3 to 8, ILD 2 = LM 1 to 7, ILD 3 = LM 7 to 4, ILD 4 = LM 1 to 4, ILD 5 = LM 2 to 6, ILD 6 = LM 2 to 4, ILD 7 = LM 3 to 5, ILD 8 = LM 1 to 3, ILD 9 = LM. 3 to 4.

The intra- and inter-observer error was evaluated for each ILD by calculating the Technical Error of Measurement (TEM) and the percentage TEM (%TEM) per modality ('dry bones', 'dry' CT models and 'clinical' CT models) [27]. In accordance with previous forensic anthropological literature, acceptable values for TEM were set at <2 mm [28] and for %TEM at <1.5% for intra-observer error and <2% for inter-observer error [29].

2.3. Part C: practical implications within forensic anthropology

Up until this point the focus has been to individually explore the two types of error inherent in the use of virtual models. However, this is not realistic for routine practice. Therefore, the practical implication was determined based on the combination of the two errors combined.

First, any ILD that had unacceptable results in TEM and %TEM, in the 'dry bone' measurements in Part B, were excluded from this analysis; their unreliability could skew the results. Second, the percentage difference was calculated for the ILDs between 'dry bones' and 'dry' CT models, and 'dry bones' and 'clinical' CT models. To do this, the mean ILDs were calculated for each of the predefined ILDs based on the measurements of the total sample (n = 27) as performed by observer 1, per modality ('dry bones', 'dry' CT models, and 'clinical' CT models). These modality specific mean ILDs were used to calculate the percentage difference between the 'dry bone' measurements and 'dry' CT models, and between the 'dry bone' measurements and 'clinical' CT models, for each ILD respectively. Third, the

percentage difference was calculated for ILDs between observers for 'dry bone' measurements only. To do this, a mean ILD per observer was calculated for each of the predefined ILDs as performed by observer 1 (round 1) and observer 2 respectively. These observer specific mean ILDs were used to interpret the percentage difference between modalities for each ILD.

The percentage difference between modalities was considered acceptable if the percentage difference did not exceed the percentage difference between observers for the 'dry bone' measurements only.

3. Results

3.1. Part A: accuracy of 3D virtual modeling

Based on the calculated arithmetic mean for the overall os coxae, the surface of the 'dry' CT models deviated, on average, −0.36 mm to +0.45 mm from the 'optical 3D' models. The minimum and maximum deviations of the models were −0.27 mm to +2.86 mm (Table 2 and Fig. 2). The surface of the 'clinical' CT models deviated, on average, 0.64 mm to 0.88 mm from the 'optical 3D' models. The minimum and maximum deviations of the models were −4.99 mm to +5.00 mm (Table 2 and Fig. 2).

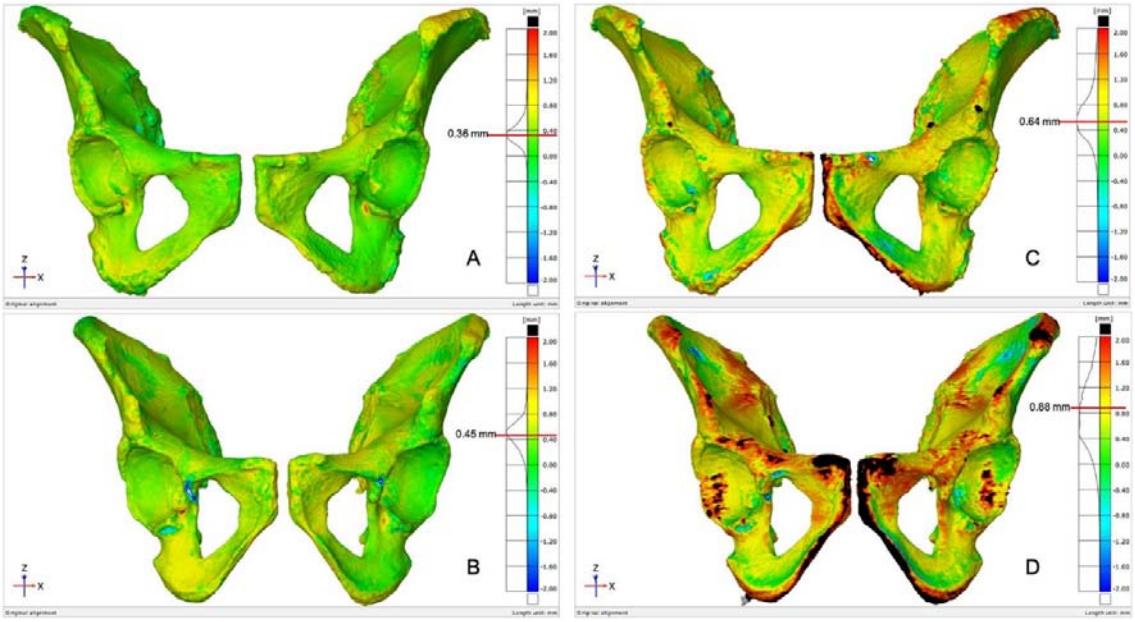

Fig. 2. Example heat maps showing the overall deviation (represented by the arithmetic mean and indicated with a red line on the histogram) in model surfaces between the segmented bone model and its optically scanned reference model. The colors represent varying degrees of accuracy: < −2 mm (white), −2 mm (dark blue), 0 mm (green), 2 mm (red) and > 2 mm (black).

Table 2. Overall size differences between 3D virtual bone models ('dry' CT models and 'clinical' CT models). These 3D virtual bone models are compared to the actual dry bones, as represented by the 'optical 3D' models. (Left and right os coxae combined).

Pelvis number	'Dry' CT models versus 'Optical 3D' models			'Clinical' CT models versus 'Optical 3D' models		
	Minimum deviation (mm)	Maximum deviation (mm)	Arithmetic mean (mm)	Minimum deviation (mm)	Maximum deviation (mm)	Arithmetic mean (mm)
1	−1.87	2.64	0.37	−4.99	5.00	0.70
2	−1.18	2.28	0.36	−3.18	4.79	0.64
3	−1.32	2.86	0.40	−4.70	4.98	0.85
4	−2.87	2.33	0.45	−3.82	4.99	0.88
5	−0.80	1.86	0.40	−2.62	4.79	0.70

Left: The smallest (A) and largest (B) deviation when comparing surfaces of the 'dry' CT models to the actual bone as represented by the 'optical 3D' model. Right: The smallest (C) and largest (D) deviation of the overall os coxa when comparing surfaces of the 'clinical' CT models to the actual bone as represented by the 'optical 3D' model.

As can be seen in Fig. 3, Fig. 4, the arithmetic means for the ROIs associated with each LM varied per ROI. The arithmetic means were consistently smaller for 'dry' CT models than for the 'clinical' CT models, with the exception of the LM 5 ROI. LM 3 showed the highest mean deviation for the ROI when comparing both the 'dry'- and 'clinical' CT models to the 'optical 3D' models.

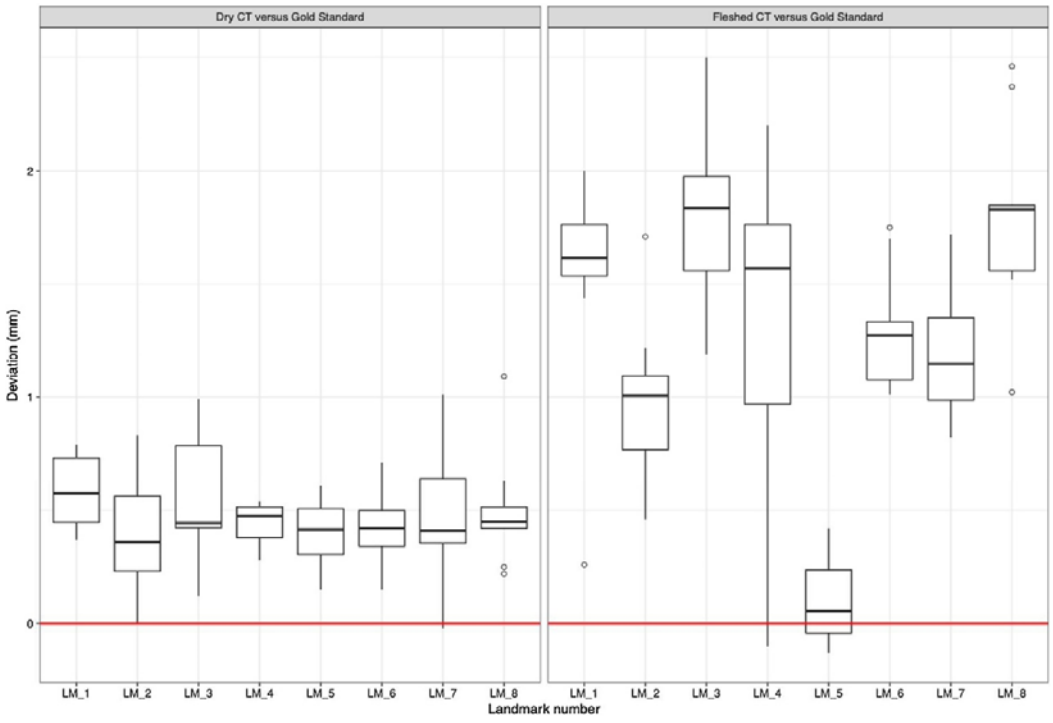

Fig. 3. Boxplot showing the variation (minimum, median and maximum) per region of interest (ROI) associated with the various landmarks (LM 1–8 as defined in Table 1), when comparing the 'dry' CT models (left) and the 'clinical' CT models (right) to the 'optical 3D' models. (Results of five os coxae, left and right combined).

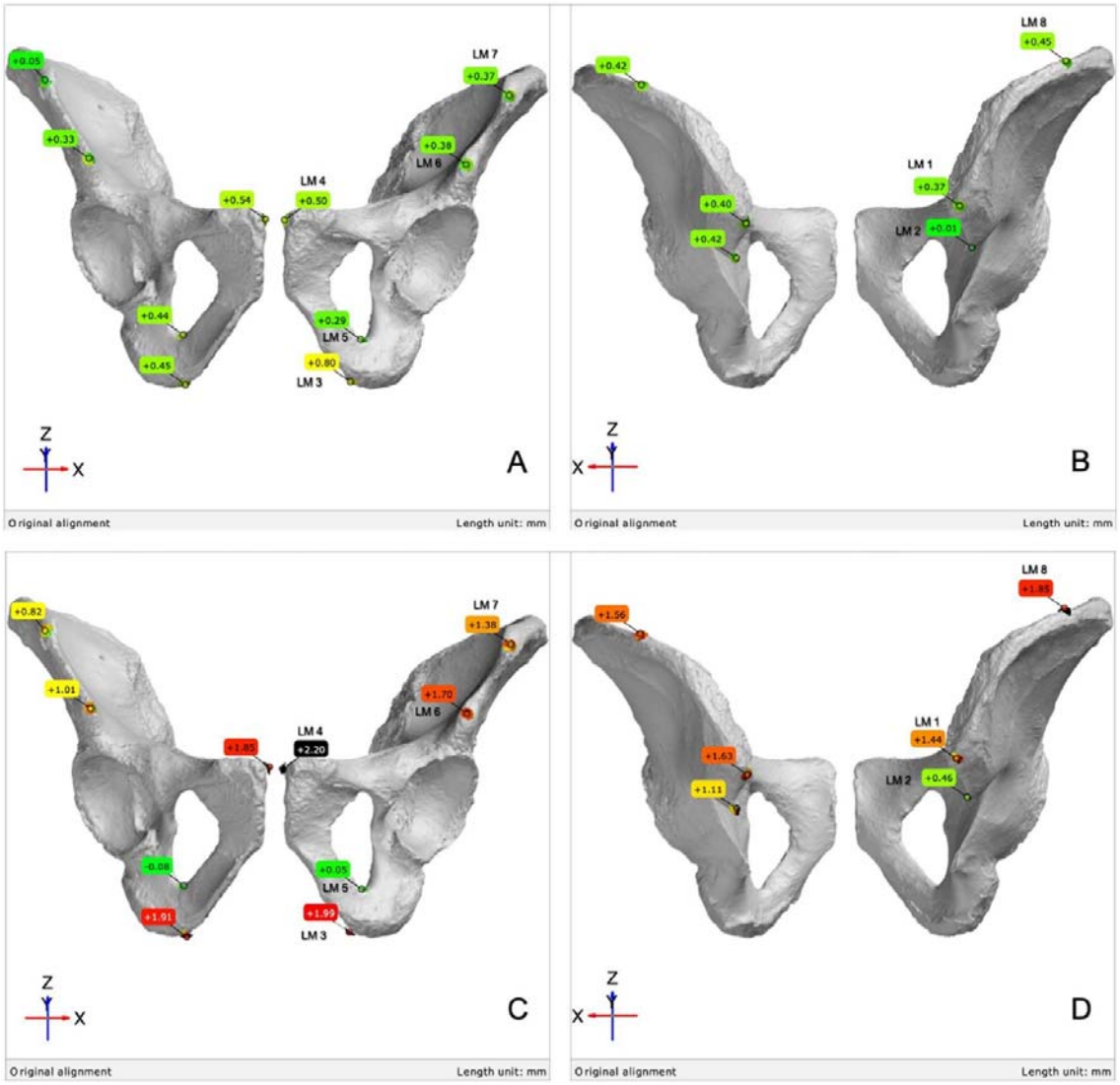

Fig. 4. Example diagram of one os coxae showing the regions of interest associated to the various landmarks (LM) (as defined in Table 1), as well as the variation between the surfaces when comparing the 'dry' CT models ((A) anterior view and (B) posterior view) and 'clinical' CT models ((C) anterior view left and (D) posterior view) to the actual bone as represented by the 'optical 3D' model.

3.2. Part B: landmark recognition

The intra-observer errors (TEM and %TEM) calculated for ILDs 1–6 were all within the generally acceptable ranges, irrespective of modality ('dry bones', 'dry' CT models, and 'clinical' CT models) (Fig. 5). This was not the case for the TEM of ILD 8 when measured on the 'dry bones' and 'clinical' CT models and ILD 9 when measured on the 'clinical CT' models. Additionally, the %TEM of ILD 7 when measured on 'dry bones' and 'clinical' CT models, ILD 8 when measured on the 'dry bones', and ILD 9 when measured on 'dry' CT models and 'clinical' CT models all exceeded the acceptable levels. Despite the TEM and %TEM being within acceptable limits for the majority of ILD measurements, some differences were noted when considering the different modalities. The intra-observer error was generally larger on 'dry bones' when compared to similar measurements on both 'dry'-

and ‘clinical’ CT models (with the exception of ILD 1, 3, and 9). Additionally, the TEM and %TEM of all ILD measurements were consistently smaller on ‘dry’ CT models than on ‘clinical’ CT models. The intra-observer error was consistently lower than the inter-observer error, for all ILDs, across all modalities.

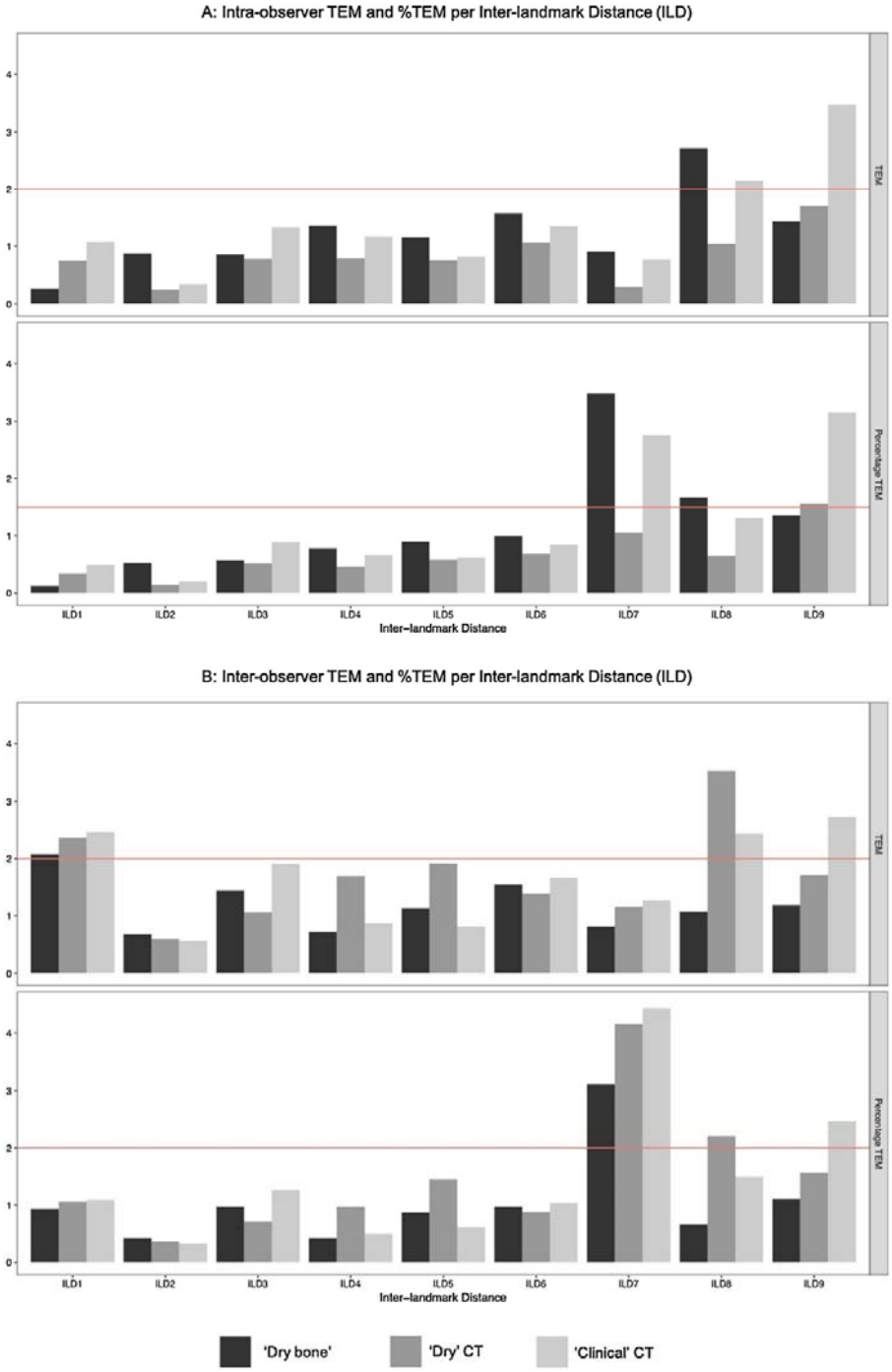

Fig. 5. TEM and %TEM for intra- observer error (A: top) and inter-observer error (B: bottom) for all inter-landmark distances, as defined in Table 1 and illustrated in Fig. 1, per modality (‘dry bones’, dry’ CT models and ‘clinical’ CT models’). The red line indicates the acceptable ranges of TEM (<2 mm for intra- and inter-observer error), and %TEM (<1.5% for intra-observer error and <2% for interobserver error) [28,29]).

The inter-observer errors (TEM and %TEM) were in the acceptable range across all modalities ('dry bones', 'dry' CT models and 'clinical' CT models) for ILD 2-6. The TEM for ILD 1 (in all modalities), ILD 8 (on 'dry'- and 'clinical' CT models) and ILD 9 (on 'clinical' CT models) were not in the acceptable range. The %TEM of ILD 7 (across all modalities), ILD 8 (on 'dry' CT models), and ILD 9 (on 'clinical' CT models) were also beyond the acceptable limits. In contrast to the intra-observer error, the inter-observer error on 'dry bone' was only larger for ILD 2 when compared to measurements performed on both the 'dry'- and the 'clinical' CT models. Furthermore, the trend of lower TEMs and %TEMs on 'dry' CT models when compared to 'clinical' CT models was not as distinct.

Overall, there seemed to be no trend or pattern with more or less reliability being associated with any single modality. Furthermore, it is important to note that the ILDs with the highest intra- and inter-observer error (ILD 1, 7, 8 and 9), all included LM 3, the ischial tuberosity.

3.3. Part C: practical implications within forensic anthropology

When comparing measurements collected on 'dry bone' to measurements collected on the 'dry' CT model, the average percentage difference for ILDs (2–6, and 9) combined was −0.90%, with a range from −0.39% (ILD 5) and −2.72% (ILD 9) (Table 3). The average percentage difference was generally larger than the average percentage difference between observers for the 'dry bone' measurements only, which was 0.37% (ranging from 0.22% to 0.91%). Exceptions to this trend are ILD 3 and 6, in which the percentage differences were smaller in this particular comparison (Table 3).

Table 3. Percent differences between the 'dry bone' measurements and the 'clinical' CT models, and between the 'dry bone' measurements versus the 'dry' CT models (for observer 1 only), for each inter-landmark distance (ILD). Additionally, the percentage difference between observers (round 1) based on the 'dry bone' measurements only, considered the standard to which the percentage difference between modalities could be compared. Only ILDs with acceptable TEM and %TEM for 'dry bone' measurements as found in Part A are included. (The ILDs are defined in Table 1 and illustrated in Fig. 1).

Inter-landmark Distances	Percentage Difference		
	[a]'Dry bones' versus 'Dry' CT models	[b]'Dry bones' Versus 'Clinical' CT models	[c]'Dry bone' between two observers
ILD 2	−0.66%	−3.20%	−0.22%
ILD 3	−0.40%	−1.72%	0.50%
ILD 4	−0.63%	−3.93%	0.22%
ILD 5	−0.39%	−1.60%	−0.06%
ILD 6	−0.59%	−1.88%	0.91%
ILD 9	−2.72%	−4.05%	0.84%
Average for all ILDs	−0.90%	−2.73%	0.37%

[a]Observer 1 ILDs on 'dry bones' versus Observer 1 ILDs on 'dry' CT models.
[b]Observer 1 ILDs on 'dry bones' versus Observer 1 ILDs on 'clinical' CT models.
[c]Observer 1 ILDs on 'dry bones' versus Observer 2 ILDs on 'dry bones'. Used as the standard to which the percentage difference between modalities could be compared.

The average percentage difference when comparing measurements collected on 'dry bone' to measurements collected on 'clinical' CT model measurements was −2.73%, with a range from −1.60% (ILD 5) and −4.05% (ILD 9). Again, the average percentage differences were larger than the average percentage difference between observers for the 'dry bone' measurements, albeit to a substantially higher magnitude than when comparing measurements from 'dry bone' with 'dry' CT models measurements (Table 3).

4. Discussion

'Clinical' CT scans prove to be consistently larger in overall size when compared to their 'dry bone' counterparts. The same holds for the 3D virtual bone models derived from the 'dry' CT scans, albeit to a consistently lesser extent. On average, the arithmetic mean difference was less than 1 mm, however, specific landmarks could have up to a 5 mm difference. The largest differences in size in both modalities was largely focused on the anteromedial pubic bone, specifically the ischio-pubic ramus, the pubic symphysis, and the ischial tuberosity, though other areas, such as the iliac crest were also problematic. Any ROI in the areas that displayed the largest differences between modalities resulted in a subsequently greater error in measurements. The observed differences between the modalities could have significant practical implications when constructing ILDs, since ILDs taken from ROIs with large (>1 mm) variability can result in a measurement error larger than the generally accepted 2 mm error. In general, ROIs on 'dry' CT models showed variation <1 mm, while ROIs on 'clinical' CT models showed variation >1 mm, except for LM 5 ('most posterior point of the obturator foramen'), which showed variation smaller than that of 'dry' CT models.

In terms of reliability of each LM, results from this study confirm that the error rates of almost all ILDs (ILD 2–6) were acceptable, however, the precision with which LMs can be recognized or identified differs per LM and per modality. This is because each LM suffers from its own combination of errors (i.e. the ability/inability to manually palpate the ROI, the reconstruction error during the 3D segmentation process, and the measurement protocols performed on different modalities). Discerning which error(s) influence(s) each LM proved to be complex and all in all, it was impossible to come to a general conclusion of whether one modality is more or less precise (in terms of intra- and inter-observer error) with regard to LM recognition or identification. The question of whether any of the studied ILDs in the current study can in fact be used interchangeably between the different modalities could only be answered positively for ILD 3 and 6, between measurements on 'dry bone' versus from 'dry' CT virtual models. None of the ILDs presented in this study can be applied interchangeably between 'dry bone' and 'clinical' CT.

Percentage differences of measurements between modalities far exceeded the percentage difference between observers for dry bone measurements. This means that the variation in size found among measurements taken from dry bones and 3D virtual bone models are not only influenced by measurement- and/or landmark recognition error, but that the modality in which the measurements are taken plays a significant role. This is an expected result, given the size differences found in the first part of this study. It however is in contrast to statements made by other authors that previously attempted to answer a similar question but used postmortem CT scans to do so. For example, Stull et al. 2014 [10], mainly ascribed the variation between measurements from different modalities to errors in measurement

repeatability. Based on the current study, it is reasonable to see how they came to this conclusion considering they used post-mortem CT images, which are less affected by image noise and thus have a lower segmentation error.

In summary, our study results show that the presence of soft tissue is a critical factor in the accuracy of virtual modalities. Additionally, the variation observed between ILDs in different modalities are ILD-specific. This means that interchangeability between dry bones and 3D virtual bone models generated from clinical CT data cannot be assumed, and that to develop or test forensic anthropological estimation methods, the accuracy of every measurement of interest needs to be established for all modalities. Additionally, when the differences between the modalities are too large, the modality on which the method was developed should be the only modality to which the method is further applied.

4.1. Influence of soft tissue presence and imaging parameters on accuracy

The differences in size found between the dry bones and the 'clinical' virtual models generated from clinical CT data are most likely associated with image noise produced by the presence of soft tissue. Attempts by previous studies to describe the accuracy of 3D virtual bone models also did so by measuring predefined ILDs [10,11], and reported little to no differences between the actual dry bones and their 3D virtual model counterparts. Unfortunately, with these studies, there were either small sample sizes or they used dry/skeletonized- or partially decomposing remains [11,16], and/or higher radiation dosages [10] (480 mA) [30], which explains why there were minimal size differences between objects. CT scans of a skeletonized bone (as performed by for example Franklin et al. (2013) [11]) or a single leg (as was performed by Robinson et al. (2008) [16]) would be scanned with a higher dose of radiation than a CT scan of whole body, therefore resulting in improved image quality. This especially holds true for reconstructions of the pelvis derived from a total body CT scan; the pelvis is known for having high levels of attenuation and image noise [17].

Additionally, the influence of radiation dosage was previously studied by Oka et al. (2009) [14] using CT scans of 12 dry forearm bones, which were acquired with two different radiation doses (50 mA and 10 mA). The 3D reconstructions of both sets of CT scans derived from both radiation doses showed the same level of accuracy. This could be explained by the lack of soft tissue present and the use of a single (non-complex) bone. It is known that the segmentation process of bones with complex structures and joint surfaces is more difficult and would thus influence the accuracy of the virtual bone model [21].

The results of previous studies are therefore not representative for 3D reconstructions based on clinical data.

4.2. Factors influencing landmark recognition

Three possible factors may have influenced the precision with which some LMs could be recognized or identified in the various modalities: the ability/inability to manually palpate the ROI, the reconstruction error during the 3D segmentation process, and the measurement protocols performed on different modalities.

Firstly, the ability to manually palpate a LM is considered to increase its identification [10,11,31]. This may suggest that LM recognition would potentially be more precise on 'dry bones' than on 3D models. However, the current study shows that the precision differed per LM and modality, which suggests that a possible improvement of manual palpation is LM specific. For instance, the intra- and interobserver error for ILD 2 were highest for 'dry bone' and lowest for 'dry' CT and 'clinical' CT respectively.

Secondly, the quality of the segmentation will influence the precision of virtual modelling and consequently detection of the LMs of interest. Based on previously published literature [21], anatomical regions such as joint surfaces and areas typically associated with enthesophyte development are less precise. However, in this study, the precision of the LM did not necessarily influence the ability to identify the LM. This is demonstrated by LM 3 ('farthest point of ischial curve from the center of the obturator foramen'). The geometrical region associated with this landmark has proven to be an area with high precision [21]. Despite the high precision of the landmark, four inter-landmark distances that could not be measured reliably on one or more of the modalities included LM 3. Error associated with this landmark is therefore not an artefact of the reconstruction process but rather stems from the ability to correctly identify this specific landmark.

Thirdly, performing measurements on dry bones is a fairly straightforward procedure of manually locating the landmark and measuring the ILD with a caliper. Measurement protocols on 3D virtual bones are often more complicated, which may also influence the precision of measurements. Previous studies have used a combination of line tools and cross-sectional views [11] or transparent bone algorithms [16] to be able to locate landmarks on 3D reconstructions. However, the same measurement protocol may not always be used for two different measurements of the same bone (as indicated by Franklin et al. (2013) [11]). It is therefore imperative that the measurement protocol has clear descriptions on how landmarks should be located and measured in order for the resulting measurements to be reliable across observers. To overcome the possible influence of different measuring protocols and the associated errors, the presented study treated and measured the 3D virtual bones as if they were 'dry bones'. This means that the measurement protocols of the dry bones and 3D virtual bones were identical and no cross-sectional- or transparent views were used to locate the landmarks. The authors thus feel that the errors presented are representative of the precision with which landmarks can be identified across modalities, rather than due to the differences in measuring protocols.

4.3. Limitations of the study

The use of a single segmentation method may be considered a limitation, as different segmentation methods may result in slightly different virtual bone models [32]. Therefore, the variability between segmentation methods, per landmark, should be tested before methods are used interchangeably between the actual dry bones and 3D virtual bone models.

Despite the ROIs being manually selected and thus introducing a potential for intra-observer error, the x,y,z coordinates of each ROI was recorded and used to recreate the ROI on the corresponding models. This means that the selection process only took place once per

'optical 3D model', and thus the authors believe that the ROIs sufficiently represents the landmarks without including landmark recognition error. Additionally, the 4 mm sized sphere, with a radius of 2 mm from the center point of the sphere, represents the acceptable 2 mm error that is reported within traditional forensic anthropological research.

The use of only eight landmarks in this study may be considered a limitation. However, these landmarks were carefully selected: a) to represent various levels of segmentation error, and b) to include landmarks generally used in forensic anthropology estimation techniques. Therefore, the authors feel that the number of landmarks allow for useful and valid results.

The older age of individuals included in this study, and consequently the presence of enthesophytes, may have influenced landmark recognition. This is a possible explanation for the errors associated with LM 3. However, the use of older individuals represents a worse-case scenario and the reported results are considered to be a reflection hereof.

5. Conclusion

This is the largest study to date where the accuracy of virtual bone models compared to their dry bone counterparts is quantified. Previous studies have investigated accuracy in a different way, for example; using TEM and %TEM, as well as percentage difference. Both techniques include the compounded error associated to size differences and landmark recognition. The cumulative effect of these errors is partially predictable, i.e. larger error between 'dry bone' and 'clinical' CT versus 'dry bone' and 'dry CT' due to the presence of soft tissue, and also partially unpredictable, i.e. ability to precisely locate landmarks across different modalities. Each landmark and subsequent inter-landmark distance therefore results in different degrees of measurement error, attributed to both differences in size due to differences in modalities and difficulties in landmark recognition.

Results from this study show that there is no pattern regarding whether one modality is more or less reliable. This means that before forensic anthropological estimation methods can be developed from 3D virtual bone models and applied to dry skeletal remains, or vice versa, the error for both landmark recognition and size between the modalities for each measurement must be established.

CRediT authorship contribution statement

Kerri L. Colman: Investigation, Conceptualization, Writing - original draft, Methodology, Formal analysis, Visualization, Project administration. **Hans H. de Boer:** Supervision, Writing - review & editing, Conceptualization, Methodology. **Johannes G.G. Dobbe:** Software, Conceptualization, Writing - review & editing, Methodology. **Niels P.T.J. Liberton:** Software, Methodology, Resources. **Kyra E. Stull:** Conceptualization, Writing - review & editing. **Maureen van Eijnatten:** Methodology, Resources. **Geert J. Streekstra:** Methodology, Resources. **Roelof-Jan Oostra:** Supervision, Resources. **Rick R. van Rijn:** Supervision, Resources. **Alie E. van der Merwe:** Supervision, Writing - review & editing, Conceptualization, Methodology.

Acknowledgments

The authors would like to thank Aubrey van het Reve, Hannah Crijns and Roy Snijckers for their participation in this study.

References

[1] S.J. Decker, S.L. Davy-Jow, J.M. Ford, D.R. Hilbelink. **Virtual determination of sex: metric and nonmetric traits of the adult pelvis from 3D computed tomography models.** J. Forensic Sci., 56 (5) (2011), pp. 1107-1114

[2] D. Franklin, A. Flavel, A. Kuliukas, A. Cardini, M.K. Marks, C. Oxnard, P. O'Higgins. **Estimation of sex from sternal measurements in a Western Australian population.** Forensic Sci. Int., 217 (1-3) (2012) 230 e1-5

[3] D. Franklin, A. Cardini, A. Flavel, M.K. Marks. **Morphometric analysis of pelvic sexual dimorphism in a contemporary Western Australian population.** Int. J. Legal Med., 128 (5) (2014), pp. 861-872

[4] M. Djorojevic, C. Roldán, P. García-Parra, I. Alemán, M. Botella. **Morphometric sex estimation from 3D computed tomography os coxae model and its validation in skeletal remains.** Int. J. Legal Med., 128 (5) (2014), pp. 879-888

[5] S. Torimitsu, Y. Makino, H. Saitoh, A. Sakuma, N. Ishii, D. Yajima, G. Inokuchi, A. Motomura, F. Chiba, R. Yamaguchi. **Morphometric analysis of sex differences in contemporary Japanese pelves using multidetector computed tomography.** Forensic Sci. Int., 257 (2015), pp. 530.e1-530.e7

[6] A. Clavero, M. Salicrú, D. Turbón. **Sex prediction from the femur and hip bone using a sample of CT images from a Spanish population.** Int. J. Legal Med., 129 (2) (2015), pp. 373-383

[7] K.L. Colman, M.C.L. Janssen, K.E. Stull, R.R. van Rijn, R.J. Oostra, H.H. de Boer, A.E. van der Merwe. **Dutch population specific sex estimation formulae using the proximal femur.** Forensic Sci. Int., 286 (2018), pp. 268.e1-268.e8

[8] A.L. Brough, G.N. Rutty, S. Black, B. Morgan. **Post-mortem computed tomography and 3D imaging: anthropological applications for juvenile remains.** Forensic Sci. Med. Pathol., 8 (3) (2012), pp. 270-279

[9] M. Aalders, N. Adolphi, B. Daly, G. Davis, H. De Boer, S. Decker, J. Dempers, J. Ford, C. Gerrard, G. Hatch. **Research in forensic radiology and imaging; identifying the most important issues.** J. Forensic Radiol. Imaging, 8 (2017), pp. 1-8

[10] K.E. Stull, M.L. Tise, Z. Ali, D.R. Fowler. **Accuracy and reliability of measurements obtained from computed tomography 3D volume rendered images.** Forensic Sci. Int., 238 (2014), pp. 133-140

[11] D. Franklin, A. Cardini, A. Flavel, A. Kuliukas, M.K. Marks, R. Hart, C. Oxnard, P. O'Higgins. **Concordance of traditional osteometric and volume-rendered MSCT interlandmark cranial measurements.** Int. J. Legal Med., 127 (2) (2013), pp. 505-520

[12] M.A. Verhoff, F. Ramsthaler, J. Krähahn, U. Deml, R.J. Gille, S. Grabherr, M.J. Thali, K. Kreutz. **Digital forensic osteology—possibilities in cooperation with the Virtopsy® project.** Forensic Sci. Int., 174 (2-3) (2008), pp. 152-156

[13] C.F. Hildebolt, M.W. Vannier, R.H. Knapp. **Validation study of skull three-dimensional computerized tomography measurements.** Am. J. Phys. Anthropol., 82 (3) (1990), pp. 283-294

[14] K. Oka, T. Murase, H. Moritomo, A. Goto, K. Sugamoto, H. Yoshikawa. **Accuracy analysis of three-dimensional bone surface models of the forearm constructed from multidetector computed tomography data.** Int. J. Med. Robot. Comput. Assist. Surg., 5 (4) (2009), pp. 452-457

[15] P.M. Lopes, C.R. Moreira, A. Perrella, J.L. Antunes, M.G. Cavalcanti. **3-D volume rendering maxillofacial analysis of angular measurements by multislice CT.** Oral Sur. Oral Med. Oral Pathol. Oral Radiol. Endodontol., 105 (2) (2008), pp. 224-230

[16] C. Robinson, R. Eisma, B. Morgan, A. Jeffery, E.A. Graham, S. Black, G.N. Rutty. **Anthropological measurement of lower limb and foot bones using multi-detector computed tomography.** J. Forensic Sci., 53 (6) (2008), pp. 1289-1295

[17] L.W. Goldman. **Principles of CT: radiation dose and image quality.** J. Nucl. Med. Technol., 35 (4) (2007), pp. 213-225

[18] T. Kersten, M. Lindstaedt, D. Starosta. **Comparative geometrical accuracy investigations of hand-held 3d scanning systems-an update, International Archives of the Photogrammetry.** Remote Sens. Spatial Inform. Sci., 42 (2) (2018)

[19] J.G. Dobbe, S.D. Strackee, A. Schreurs, R. Jonges, B. Carelsen, J.C. Vroemen, C.A. Grimbergen, G.J. Streekstra. **Computer-assisted planning and navigation for corrective distal radius osteotomy, based on pre-and intraoperative imaging.** IEEE Trans. Biomed. Eng., 58 (1) (2011), pp. 182-190

[20] J.G. Dobbe, M.G. de Roo, J.C. Visschers, S.D. Strackee, G.J. Streekstra. **Evaluation of a quantitative method for carpal motion analysis using clinical 3D and 4D CT protocols.** IEEE Trans. Med. Imaging (2018)

[21] K.L. Colman, J.G. Dobbe, K.E. Stull, J.M. Ruijter, R.J. Oostra, R.R. van Rijn, A.E. van der Merwe, H.H. de Boer, G.J. Streekstra. **The geometrical precision of virtual bone models derived from clinical computed tomography data for forensic anthropology.** Int. J. Legal Med. (2017)

[22] **Artec Space Spider Specifications** (2018)

<https://www.artec3d.com/portable-3d-scanners/artec-spider#specifications> (Accessed 5 November 2018)

[23] L. Betti, N. von Cramon-Taubadel, A. Manica, S.J. Lycett. **Global geometric morphometric analyses of the human pelvis reveal substantial neutral population history effects, even across sexes.** PLoS One, 8 (2) (2013), Article e55909

[24] J.A. Bytheway, A.H. Ross. **A geometric morphometric approach to sex determination of the human adult os coxa.** J. Forensic Sci., 55 (4) (2010), pp. 859-864

[25] N.R. Langley, L.M. Jantz, S.D. Ousley, R.L. Jantz, G. Milner. **Data Collection Procedures for Forensic Skeletal Material 2.0.** University of Tennessee and Lincoln Memorial University (2016)

[26] K.L. Colman, J.G.G. Dobbe, K.E. Stull, J.M. Ruijter, R.J. Oostra, R.R. van Rijn, A.E. van der Merwe, H.H. de Boer, G.J. Streekstra. **The geometrical precision of virtual bone models derived from clinical computed tomography data for forensic anthropology.** Int. J. Legal Med., 131 (4) (2017), pp. 1155-1163

[27] T.A. Perini, G.Ld. Oliveira, Jd.S. Ornellas, F.Pd. Oliveira. **Technical error of measurement in anthropometry.** Rev. Bras. Med. Do Esporte, 11 (1) (2005), pp. 81-85

[28] K.E. Stull, E.N. L'abbé, S. Steiner. **Measuring distortion of skeletal elements in L odox S tatscan-generated images.** Clin. Anat., 26 (6) (2013), pp. 780-786

[29] N.R. Langley, L.M. Jantz, S. McNulty, H. Maijanen, S.D. Ousley, R.L. Jantz. **Error quantification of osteometric data in forensic anthropology.** Forensic Sci. Int., 287 (2018), pp. 183-189

[30] K.E. Stull, M.L. Tise, Z. Ali, D.R. Fowler. **Accuracy and reliability of measurements obtained from computed tomography 3D volume rendered images.** Forensic Sci. Int., 238 (2019), pp. 133-140 (personal communication)

[31] K. Aldridge, S.A. Boyadjiev, G.T. Capone, V.B. DeLeon, J.T. Richtsmeier. **Precision and error of three-dimensional phenotypic measures acquired from 3dMD photogram-metric images.** Am. J. Med. Genet. A, 138 (3) (2005), pp. 247-253

[32] M. van Eijnatten, R. van Dijk, J. Dobbe, G. Streekstra, J. Koivisto, J. Wolff. **CT image segmentation methods for bone used in medical additive manufacturing.** Med. Eng. Phys., 51 (2018), pp. 6-16