Elsevier Editorial System(tm) for Forensic Science International

Manuscript Draft

Manuscript Number:

Title: Facial Soft Tissue Thicknesses in Craniofacial Identification: Data Collection Protocols and Associated Measurement Errors

Article Type: Review Article

Keywords: Craniofacial identification; facial approximation; facial reconstruction; craniofacial superimposition; facial soft tissue depth; skull; measurement error

Corresponding Author: Dr. Carl Stephan, PhD

Corresponding Author's Institution: The University of Queensland

First Author: Carl Stephan, PhD

Order of Authors: Carl Stephan, PhD; Brandon Meikle; Nicolle Freudenstein; Ronn Taylor; Peter Claes

Abstract: Facial soft tissue thicknesses (FSTT) form a key component of craniofacial identification methods, but as for any data, embedded measurement errors are highly pertinent. These in part dictate the effective resolution of the measurements. As herein reviewed, measurement methods are highly varied in FSTT studies and associated measurement errors have generally not been paid much attention. Less than half (44%) of 95 FSTT studies comment on measurement error and not all of these provide specific quantification. Where informative error measurement protocols are employed (5% of studies), the mean error magnitudes range from 3% to 45% rTEM and are typically in the order of 10-20%. These values demonstrate that FSTT measurement errors are similar in size to (and likely larger than) the magnitudes of many biological effects being chased. As a result, the attribution of small millimeter or submillimeter differences in FSTT to biological variables must be undertaken with caution, especially where they have not been repeated across different studies/samples. To improve the integrity of FSTT studies and the reporting of FSTT measurement errors, we propose the following standard: (1) calculate the technical error of measurement (TEM or rTEM) in any FSTT research work; (2) assess the error embedded in the full data collection procedure; and (3) conduct validation testing of FSTT means proposed for point estimation prior to publication to ensure newly calculated FSTT means provide improvements. In order to facilitate the latter, a freely available R tool TDValidator that uses the C-Table data for validation testing is herein provided.

Suggested Reviewers: Ginesse Listi PhD
Louisiana State University
glisti1@lsu.edu
Experienced person in the field and President of the International Association of Craniofacial Identification. Also, Dr Listi is an American Board Certified forensic anthropologist who is active in casework.

Pierre Guyomarc'h PhD

ICRC
pierreguyo@gmail.com
Dr Guyomarc'h has published previously on facial soft tissue thicknesses derived by CT scan and is a manager within the ICRC.

Maryna Steyn PhD
The University of Witwatersrand
Maryna.Steyn@wits.ac.za
Prof. Steyn has published multiple prior papers on facial soft tissue thicknesses, is head of department at the University of Wits, and is former president of the International Association for Craniofacial Identification.

Opposed Reviewers: Caroline Wilkinson
C.M.Wilkinson@ljmu.ac.uk
Personal conflict. She does not like some members of this author team.

Facial Soft Tissue Thicknesses in Craniofacial Identification: Data Collection Protocols and Associated Measurement Errors*

C.N. Stephan[1], B. Meikle[1], N. Freudenstein[2], R. Taylor[1], P. Claes[3,4,5]

[1] The Laboratory for Human Craniofacial and Skeletal Identification (HuCS-ID Lab), School of Biomedical Sciences, The University of Queensland, St Lucia, Australia, 4072.

[2] Institute for Forensic Medicine, University of Leipzig, Leipzig, 04103, Germany.

[3] Department of Electrical Engineering (ESAT) / Processing of Speech and Images (PSI), KU Leuven, Leuven, Belgium.

[4] Medical Imaging Research Center (MIRC), UZ Gasthuisberg Leuven, Leuven, Belgium.

[5] Department of Human Genetics, KU Leuven, Leuven, Belgium.

***** The views and opinions expressed herein are entirely those of the authors. They are not to be construed as official views of any institutions, editorial boards, or governing boards to which the authors may be affiliated.

Running Head: Measurement Error in Facial Tissue Thicknesses

Corresponding author:

Carl N. Stephan

Laboratory for Human Craniofacial and Skeletal Identification

School of Biomedical Sciences, The University of Queensland, Brisbane, Australia, 4072

Ph: +61 7 3365 7485

Email: c.stephan@uq.edu.au

Highlights

- Facial soft tissue thickness are a cornerstone of craniofacial identification
- Many methods have been used to measure facial soft tissue thicknesses
- Measurement errors appear to be large and have not been sufficiently documented
- We review pre-existing data and recommend new future standards

Abstract

Facial soft tissue thicknesses (FSTT) form a key component of craniofacial identification methods, but as for any data, embedded measurement errors are highly pertinent. These in part dictate the effective resolution of the measurements. As herein reviewed, measurement methods are highly varied in FSTT studies and associated measurement errors have generally not been paid much attention. Less than half (44%) of 95 FSTT studies comment on measurement error and not all of these provide specific quantification. Where informative error measurement protocols are employed (5% of studies), the mean error magnitudes range from 3% to 45% rTEM and are typically in the order of 10-20%. These values demonstrate that FSTT measurement errors are similar in size to (and likely larger than) the magnitudes of many biological effects being chased. As a result, the attribution of small millimeter or submillimeter differences in FSTT to biological variables must be undertaken with caution, especially where they have not been repeated across different studies/samples. To improve the integrity of FSTT studies and the reporting of FSTT measurement errors, we propose the following standard: (1) calculate the technical error of measurement (TEM or rTEM) in any FSTT research work; (2) assess the error embedded in the full data collection procedure; and (3) conduct validation testing of FSTT means proposed for point estimation prior to publication to ensure newly calculated FSTT means provide improvements. In order to facilitate the latter, a freely available R tool *TDValidator* that uses the C-Table data for validation testing is herein provided.

Keywords: Craniofacial identification; facial approximation; facial reconstruction; craniofacial superimposition; facial soft tissue depth; skull; measurement error

Introduction

Facial soft tissue thicknesses (FSTT) form a core component of the craniofacial identification methods helping to provide practitioners with metric guidance as to the amount of soft tissue that overlies the skull [1]. Since 1883, there have been over 95 FSTT studies published in the literature collectively tallying >246,500 tissue thickness measurements of >16,500 individuals [2]. In the last 30 years, the number of published FSTT studies has risen dramatically (Fig. 1). Almost all studies follow the same principles established as early as 1883 by Welcker [3] whereby tissue thicknesses are measured, means are calculated and results are published with little or next to no validation testing to document estimation accuracies [4]. The surging popularity of these FSTT studies makes a comprehensive review of FSTT in the context of their measurement errors timely.

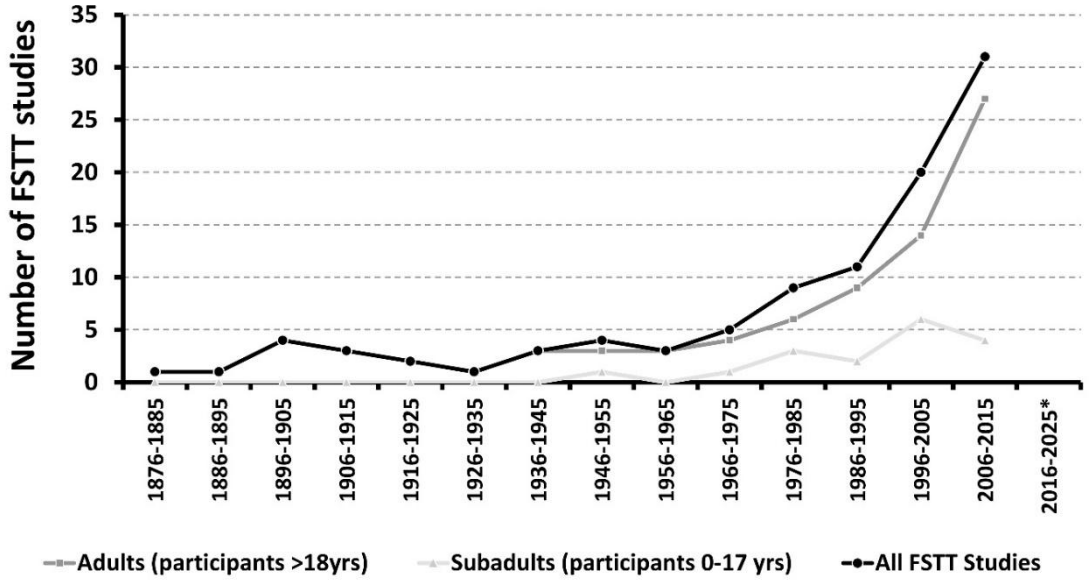

Fig. 1: FSTT studies from 1883 to 2015 (n = 98). Asterisk (*) indicates current period for which studies are not yet complete. Figure reproduced from [5] with permission from Elsevier.

With some exceptions in North America [6-9], almost all FSTT data have been used as *general guides* as to the soft tissue volume covering the skull in craniofacial identification, not exact point-estimation values [10-17]. In other words, the means are not used as strict indicators. Instead, some divergence

from the mean is tolerated when the measurements are used in practice for specific individuals [18]. This aligns with statistical principles where the mean value represents the central tendency, but not necessarily the precise value held by any one individual. Subsequently, in many facial approximation methods the FSTT means are often modified by small amounts according to the bony relief (size and shape) of the skull when used in casework [10-12]. This step is not without its own considerations / controversy since modifications may be made on practitioner's speculative whims rather than empirically documented biological relationships [18]. Irrespectively, the mean values serve as a guide not an exact measurement and in similar fashion for craniofacial superimposition methods, intervals around the mean rather than the mean measurement itself, may serve as the referent [16].

In contrast to the casework where FSTT means are used as general indicators, FSTT researchers often analyse mean FSTTs differences between groups (e.g., by sex, age, ancestry) down to the millimetre or submillimetre level of precision [19]. It is thereby pertinent to ask what the utility of such small research differences are when: 1) they are not used in practice; and 2) the measurement errors are likely as large. Data improvements can only be guaranteed by fine precision when: 1) measurement errors are small enough not to interfere with the detection of any differences, i.e., they do not bury the signal in noise [20]; and 2) improved estimation accuracies are demonstrated in practice by validation testing of the estimation models.

When noisy data are encountered, one approach to extract meaningful results is the use of long run pooled means to average out study specific errors [1, 21]. However, obtaining FSTT data with small measurement errors is in any case an advantage because better ground truth estimates can be achieved and with smaller samples. In these instances, it is vital that the measurement error be small since they will be manifested at full intensity without any dampening or buffering provided by other data. Subsequently, in either approach of single studies or pooled studies, there is a double-edged sword of error risk that must be weighed and mitigated.

So far, sufficient attention has not been paid to measurement errors in FSTT research [1, 4, 20]. Not surprisingly then there has not been a systematic review of FSTT measurement errors so far published. Much speculation has been made as to what factors *might* be important for FSTT measurement (e.g., cadaver measurements compared to living subjects [14, 19, 22-27]), but quantitative data on these factors are generally absent and the error elucidated by repeated measurement commonly goes entirely unaddressed [20]. There are, for example, no metric data in the FSTT literature that specifically demonstrate measurements from non-embalmed recently deceased individuals to be metrically inferior to data derived from living subjects despite a common perspective that this is the case [28]. To the exact contrary, there are data that instead contradict the long held speculations that these data are different, i.e., that after a period of curing, embalmed cadavers hold very similar mean values to living subjects measured by ultrasound [1] (Fig. 2 and 3). In some cases the measurement method provides restrictions for establishing the measurement error since the measurement protocol does not permit remeasurement of the same subjects—such as the case for radiographic protocols where repeat imaging sessions are universally avoided due to additional ionizing radiation doses, see e.g., [29-55]. This is a major limitation.

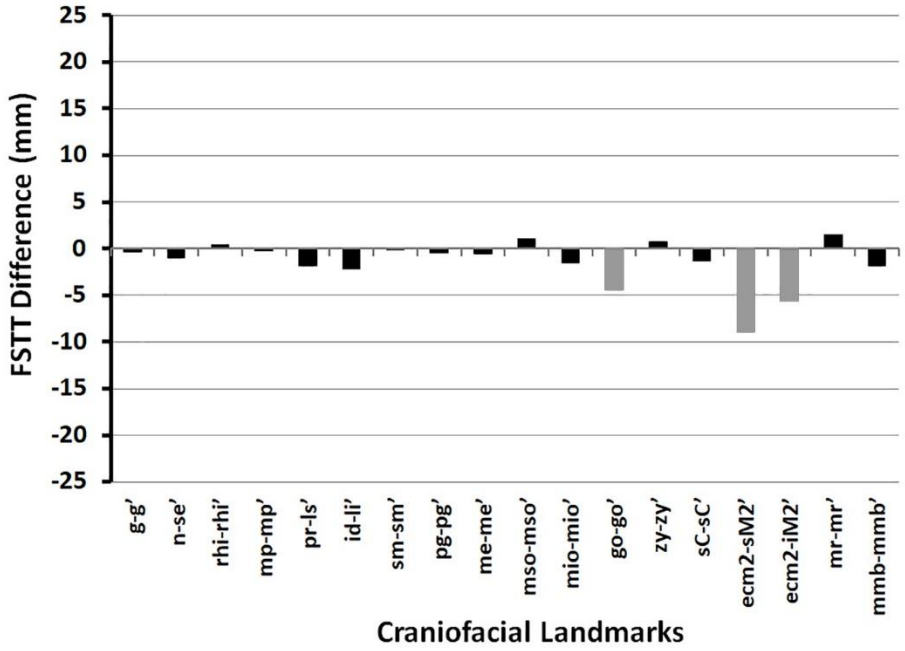

Fig. 2: FSTT differences between pooled independent cross-sectional samples of female White European cadavers measured by needle puncture (13 pooled studies N = 18-277, [56-67] and [68] cited in [69]) and living subjects measured by ultrasound (3 pooled studies, N = 505-654, [16, 23, 70]) after [1]. Negative sign indicates ultrasound > needle puncture. Note comparable data (tiny differences) at most landmarks (15 of 18 in black), and with exceptions only at three highlighted in grey (go-go′, ecm^2-$sM^{2\prime}$, ecm_2-$iM_2\prime$).

Given that the living versus deceased status of participants is a popular topic of discussion in FSTT data accuracy [14, 19, 22-27], it is important to note that challenges to accurately measure the FSTT in cadavers also equally applies to living subjects. Subsequently, these limits do not *a priori* establish measurements on living subjects as accurate or free from the impacts of measurement error. For example, the facial soft tissues are highly mobile, delicate and readily compressible in living subjects complicating measurement methods that require skin contact [16]. These tissue properties also make body posture an important consideration for non-contact measurement methods, like CT and MRI, where, to date, subjects have almost universally been measured laying down in the supine position [71-75] (Fig. 4).

Contrary to popular thought then [76], measurement of living subjects does not guarantee accuracy, and neither does use of advanced imaging technologies even if technical precision of the imaging equipment is high [77]. It is vital to recognize that equipment precision is not the same as hands-on clinical measurement of humans since the former fails to include error of practitioners taking measurements on living subjects in real world clinical settings. Subsequently, manufacturer reported equipment precision is often much higher (error lower) than that obtained in real-life clinical context where real human subjects (whose soft tissues are highly deformable) are physically measured. This is evidenced by discrepancies in results between different medical imaging methods that all are regarded 'accurate', e.g., measurements made by A-mode ultrasound are not identical to MRI data [16] and neither are ultrasound and CT data exactly the same [73]. Medical imaging techniques present yet further hurdles since obtaining repeat scans of healthy but randomly selected subjects is rare—most clinical samples are not random draws from the larger population [20, 22].

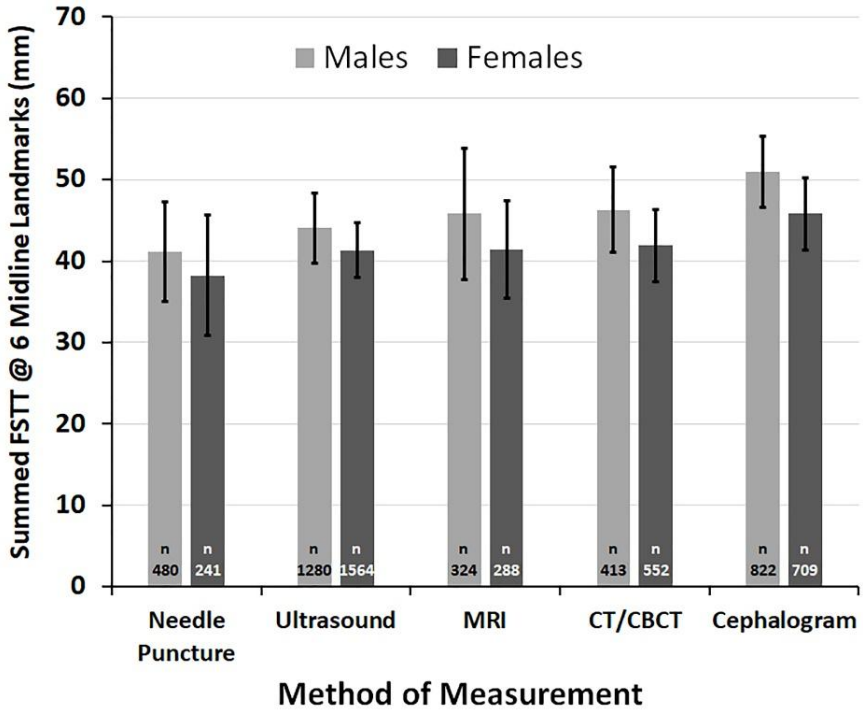

Fig. 3: Grand mean differences between different FSST measurement methods revealed by data pooling. The multivariate data for six landmarks (g-g′, n-se′, rhi-rhi′, sm-sm′, pg-pg′, me-me′) have

been simplified into a single generic indicator of the tissue volume (summed FSTT measurement). Only those studies reporting complete data for adults (>18 years) at the aforementioned six landmarks are presented as drawn from the full-suite of FSTT literature from 1883-2018. Exact studies from which this plot is generated are provided in Table 1. Bars represent ±1 standard deviation of individual study tallies around the grand mean. Note the different FSTTs across the different measurement modes and the relatively much larger cephalogram values compared to other methods as previously described in the literature [1]. While males consistently display larger raw summed values for all methods, it must be recognized that the data have not been adjusted for body size, a step that reverses the direction of the differences, i.e., females hold larger values than males [78, 79].

Fig. 4: Mean change in face shape from upright to supine posture mapped on the average face for the same subjects (n = 62). Light pixels indicate soft tissue extrusion sites in the supine position, while the dark pixels indicate retrusion of tissues in the supine position. Images reproduced from [71] with permission from Elsevier.

<Table 1 about here>

Subsequently, FSTT measurement is, in practice, a much more challenging task than what might be anticipated on first impression, especially if millimetre accuracy/precision is sought. This raises the question if small differences observed between different studies / authors / samples that are commonly emphasized in the recent FSTT literature are trustworthy, or important? This is especially pertinent to studies of FSTT by ancestry in small sampled single investigations [4]. In this paper, we explicitly review and analyse FSTT measurement protocols and errors from the prior published literature to provide a better handle on this issue. We conclude with suggestions to improve error documentation in future work and provide a new computer tool to facilitate user-friendly out-of-group validation testing of FSTT means as vital to documentation of practical utility.

FSTT Measurement Methods

Any discussion of FSTT measurement errors must begin with a description of the protocols and equipment used for measurement since these protocols, combined with the investigator technique of implementation, set the overarching error. So far in the literature there have been eight broad classes of methods used to measure FSTTs; however, within each major class there are many variations as outlined below. Six of the major methods are mainstream: solid-core needle puncture, ultrasound (A- and B-mode), 'plain film' radiographs; computed tomography (CT), cone beam computed tomography (CBCT), and magnetic resonance imaging (MRI). The two other methods are much less common, more niche and have been used in a supplementary capacity: tissue cylinder biopsy, and methods that combined two or more imaging protocols (one for the skull and one for the face).

Solid-core Needle Puncture:

Synopsis:

At designated anatomical landmarks on a cadaver, a solid-core needle with a small-calibre is inserted into the soft tissue until it reaches the bone. The depth of the needle penetration, from the skin surface, is then recorded (Fig. 5).

Variations of Technique:

1. In an attempt to avoid tissue dimpling upon needle insertion (Fig. 5), some practitioners lubricate needles prior to insertion with a small amount of oil [67, 107]. Elastic recoil of tissues, to their original starting state, can also be encouraged by gentle movement of the needle in and out of the tissue to encourage rebound [68] (Fig. 5).
2. A variety of different needles/pins have been used, from calibrated pins etched with a metric scale (such as the GPM® Skin Thickness Measuring Instrument #119) to very sharp hypodermic needles [61, 63]. Regular sewing pins [67] have also been used, as have and 0.5 mm steel wire [108] and endodontic finger pluggers [65] (Fig. 5). The diameters of the needles/pins used are relatively consistent being 0.4-0.5 mm or 26/27 gauge.

3. Some practitioners use a locator or stop on the pin/needle, such that the tissue depth remains marked on the needle and can be measured after needle extraction from the tissue. The various locators that have previously been used include: rubber stop [56, 63, 65, 66, 82, 83, 109-111]; cork stop [61]; and brass sheath [58]. Some authors have used a standard weight on top of the stop to hold the stopper at the skin surface, e.g., Sutton [61] reports using a 4 g weight on the cork plunger. Other practitioners have avoided stops entirely by using sooted needles where the soot line (wiped clean by the soft tissue) marks the needle penetration depth without the need for any externally applied pressure [57, 59, 67, 100, 102, 103].
4. Some practitioners have palpated landmarks prior to tissue depth measurement [63], while others have not. Palpation may assist locating the bony landmarks more accurately, but it may also cause soft tissue depression in advance of measurement. Intervals to allow tissues to rebound following palpation have not been considered, but may be important depending upon tissue factors such as tissue temperature at the time of measurement (e.g., when bodies have been stored in fridges [63]) and embalming status.
5. Needle puncture methods have been used on both unembalmed, e.g., [60, 63, 65, 66], and embalmed cadavers, e.g., [61, 65, 67], and the timeframes from death to measurement have been highly varied.

For un-embalmed cadavers time since death to measurement has often gone undocumented or very generally recorded, such as "as soon as possible" [63]. Where recorded, the measurements have been taken in time windows <12 hours post-mortem [65] up to as long as up to 5 days post mortem [63]. One consideration for cadaver measurement is rigor mortis, which may appear as early as 2-4 hours after death, but onset can vary with a variety of factors, e.g., intensity of exercise immediately prior to death [112]. Clearly, FSTT measurements should be avoided while the body is in the rigor state [95], however some studies have been conducted well within typical rigor mortis periods, e.g., <12 hours post-mortem [82, 83]. Typically rigor mortis will disappear at approximately 36 hours in temperate climates [112], but it can have varied onset ranging from 12 hours to 6 days depending on the

ambient temperature. Rigor persists until the very early stages of decomposition where the bonds in the skeletal muscle filaments degenerate, releasing the rigor state.

For embalmed cadavers, protocols are likely highly varied with some designed for longer fixation times than others. Often the exact embalming protocol is not reported in FSTT studies see e.g., [67], and where details are provided they are often incomplete, see e.g., [65]. Time since death to embalm, time since embalm to measurement, embalming fluid ingredients/ratios, embalming fluid volume, administration procedure and storage environment are all variables that may potentially impact on results, though they have not been specifically investigated in terms of effect on data accuracy [65].

6. FSTT measurements have been taken at a variety of different skeletal landmarks and angles from the skull surface to the soft tissue, but generally these are not well described in the needle puncture literature [3, 56, 57, 65, 67]. Often, only the craniometric landmark is reported, although a soft tissue capulometric landmark is also required [113]. As described for a range of measurement methods, it is possible to take measurements in three ways: (a) measure between specifically defined hard and soft tissue landmarks (e.g., nasion and sellion) [96]; (b) measure at angles that bisect the estimated curved surface of the bone at specific craniometric landmarks [1, 114]; or (c) take measurements at craniometric landmarks, but in planes of particular head positions such as the Frankfurt Horizontal [50, 54].

Brief History & Recent Advances: The needle puncture method evolved from Welcker's first use of a thin blade in 1883 to measure the FSTT [3]. His [56] modified the approach to first use solid-core pins with a rubber stop and since these initial studies the method has been popularly used in a variety of formats described above, see e.g., [11, 59, 61, 62, 65, 67, 80, 82, 83, 109, 115, 116], on non-embalmed recently deceased cadavers [63, 65, 66], and embalmed cadavers [61, 65, 67], [68] cited in [15].

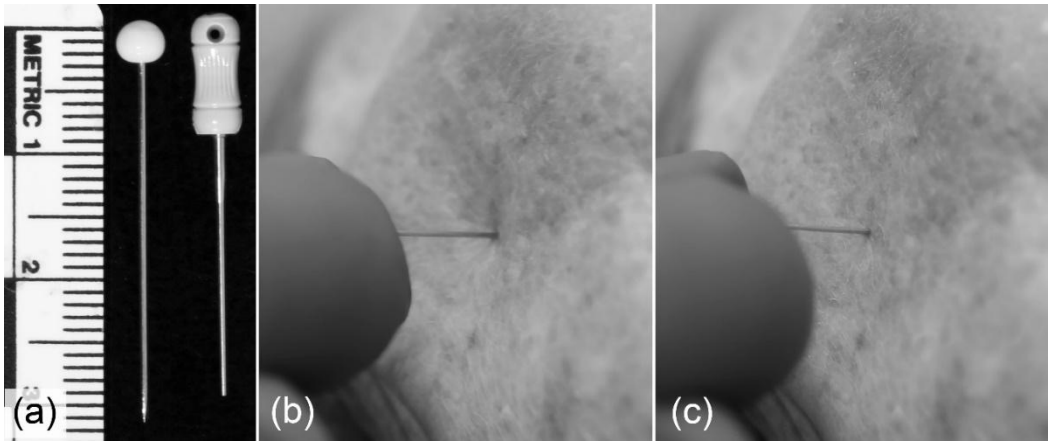

Fig 5: Needle puncture method: (a) regular polished steel sewing pin (calibre = 0.6mm, length = 30 mm, left) and SybronEndo® stainless steel Finger Plugger (size = 45, length = 21mm, right) and; (b) FSTT measurement showing a compression dimple (several millimetres deep) formed by pin point insertion of the steel sewing pin into embalmed soft tissues of the cadaver at the mid-ramus region; and (c) position for final tissue depth measurement following gentle movement of the needle back-and-forth to encourage tissue rebound. Note blunted end of the Finger Plugger (a) that for FSTT measurement requires insertion and removal of pointed pin in advance of Finger Plugger to pierce skin and provide a tract for the Finger Plugger to follow according to methods of Simpson and Henneberg [65].

Data fidelity: The needle puncture method enables direct physical measurement of the soft tissue on the subject, but it can essentially only be used on supine cadavers. As no tissue cylinder is removed to provide a clear view of the underlying skeletal surface, the position of the underlying craniometric landmarks must be entirely estimated—a major limitation [117].

No matter how sharp the needle might be, friction between the tissue and the needle causes a compression dimple to appear in the soft tissue upon contact of the needle with the skin [16]—an observation noted as early as 1909 by von Eggeling [102]. As described above, lubricated needles can limit this tissue dimpling and light movement of the needle in-and-out, during insertion, helps the tissue to relax back towards its initial starting position. This avoids direct pressure being applied to the skin such as occurs when pinching the tissue with ‘a free hand’ as recommended by Rhine and

Campbell [82] or von Eggeling [102]. Using a stopper on the needle holds the additional risk of soft tissue compression when the stopper is manually lowered to the skin surface or applies pressure to the tissue if at the tip of the needle upon insertion—a complication entirely avoided by the use of sooted needles [67].

Preferably, subjects should be measured as soon as possible after death [11]. This too applies to embalmed cadavers. If anatomical embalming methods are used (where fixative fluid administration can be large, e.g., 40L), it is vital that measurements are not taken until after a curing phase, where excess water administered in the embalming fluid has been allowed time to evaporate out of the tissues, returning them towards their original starting volumes [65]. Measurements should be avoided at any sites where soft tissue compression due to body position is observed, e.g., at the opisthocranion, if the body is stored in supine position.

While use of cadavers and embalming technique have been cause for data accuracy concern in the past [14, 19, 22-27], quantitative data show that needle puncture on cadavers produces near identical mean values to ultrasound on living subjects [1, 87, 88]. The most salient difference in the raw data between these two measurement modes appears to be the longer tails observed in the needle puncture distributions [88]. It should be noted that cadavers from anatomy school programs also rarely represent random samples that are preferred [20]. One of the great advantages of the method is that it represents a simple and direct measurement of the soft tissues whose results can be observed immediately with the naked eye and without the need for any intermediary imaging.

Tissue Cylinder Biopsy:

Synopsis:

A dermal biopsy punch is used at designated anatomical landmarks, to remove a tissue cylinder or core sample down to the bone in cadavers (Fig. 6). The tissue is then allowed to relax/rebound over a period of days to weeks [44, 118]. Following the relax period, a measurement of the thickness of the

soft tissue immediately adjacent to the landmark of interest, is made in the sample void [44, 118] (Fig. 6).

Fig. 6: Tissue cylinder biopsy method. Image illustrates FSTT being measured at metopion; some other measurement sites also visible at the right of image, e.g., supraglabella, glabella and nasion. Image reproduced from [44] with permission from Elsevier.

Variations of Technique: Varied core calibre sizes have been used (2-3 mm [44] or 4 mm [118]) as have different time frames for tissue rebound/relax: e.g., overnight [118] versus one week [44].

Brief History & Recent Advances: The method was first used in 2005 [44] for validation testing of soft tissue data obtained from computed tomography scans (CT), and has since been used for this purpose by other investigators [118]. The method has so far been used for validation purposes of other imaging methods, rather than separately deriving FSTT means to be used as point estimators in casework.

Data fidelity: Unlike solid-core needle puncture, this method enables direct visualisation and measurement of the soft tissue near the landmark of interest, which is an advantage for ensuring correct physical measurement from underlying bony landmarks [44, 118]. As the measurements are taken directly on the individual, without any intermediary mode of imaging, the method has been used as a gold standard for testing the accuracy of CT image derived data [44, 118].

Disadvantages of the method include: (1) cadavers must be used which may not be representative of living condition as per needle puncture [73]; (2) cadavers may be weight range restricted and may not represent healthy subjects (subjects are already deceased); (3) use of cadavers forces supine body posture (not representative of upright data [71]); (4) resting of head supine, during storage, normally precludes measurement of compressed tissues over the occipital bone; (5) the method cannot be used at landmarks where underlying bone sharply undulates and/or compact bone is thin/delicate; (6) this method does not permit relocation of landmarks under blinded conditions in the same subjects for error quantification; and (7) during extraction of the tissue cylinder, the soft tissues may be compressed by the manually applied pressure to cut the punch through the soft tissue [44]. Whether or not the tissues fully rebound to initial positions following initial extraction and/or with a 'relax phase' is unknown.

Ultrasound:

Synopsis:

A transducer that emits a narrow window of high-frequency sound-wave pulses is placed on the skin in this method (with a smear of echo transmitting gel to provide echo connectivity between the transducer and the patient) to record underlying tissue interfaces [119]. Some sound waves are reflected from the tissue interfaces back to the transducer and an oscilloscope records the time taken from emission to receipt [120]. The velocity of the sound waves in soft tissue is known within ± 1 %, i.e., 1540 ± 15 m/s, making distance measurements in the direction of the sound echo accurate [121]. Some ultrasound machines enable the substrate velocity and angle of incidence to be set by the user, e.g., flaw detection machines used in engineering [70], but most clinical medicine machines work on

algorithms that assume the transducer is held perpendicular to the underlying tissue interfaces [120]. For FSTT measurement, facial hair/beards are problematic for this method so subjects possessing these features are usually excluded from research samples [70, 87, 88].

There are several types of ultrasound (including M-mode and Doppler), but for soft tissue thickness measurement either A- or B-mode is applicable:

1. A-mode stands for 'amplitude' ultrasonography [121]. Here the cathode-ray tube records the time taken for the return of the echo from the tissue interface as a peak or pulse on an amplitude graph [121] (Fig. 7). FSTT investigations that have used A-mode include: [16, 19, 24, 70, 73];
2. B-mode stands for 'brightness modulation' and a spot on the cathode-ray tube correspond to the time elapsed for return of the echo and the intensity, thus producing an image closer to an anatomical sections [121] (Fig. 7). FSTT studies using B-mode include: [23, 78, 85, 87, 88, 122].

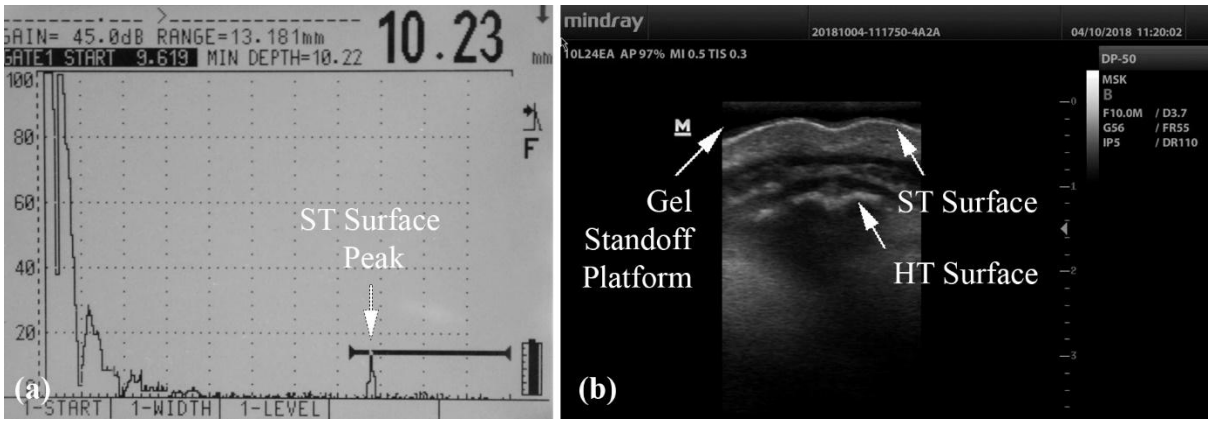

Fig. 7: Facial soft tissue thickness measurement at mid-philtrum (mp-mp′) by A-mode (a) and B-mode (b) ultrasound. The A-mode image (a) is taken using an Epoch 4b A-ultrasound device (Panametrics, Waltham, USA) identical to that used by De Greef et al. [70, 73]. The tissue distance, at the peak crossed by the black horizontal bar (gate 1), is given in the top right of the monitor, in millimetres and a pre-set tissue velocity value of 1542 m/s [73]. The B-mode image (b) is taken on a Mindray DP50 machine (Shenzen, China) with an 8.5 MHz linear transducer held in horizontal

orientation across the philtrum using a stand-off gel platform yielding a cross-section like view—note the philtrum trough and vermillion border ridges visible at the top of the image. ST = soft tissue. HT = Hard tissue.

Variations of Technique: Ultrasound transducers may be large linear devices (as in B-mode) or small circular devices (as for A-mode flaw detector machines [70, 73]). The transducer device may be placed directly against the skin with a thin smear of acoustic gel to maintain sound connectivity [73], or the unit may be placed on a thicker acoustic stand-off platform that separates the device from the skin surface (B-mode devices only) [87, 88, 123]. Liberal amounts of highly deformable acoustic gel have been used as a low viscous alternative to commercially purchased standoff gel-pads to mitigate the risk of tissue compression [87, 88].

A variety of transducer devices can be used as mentioned above and with a range of echo emission frequencies. A-mode devices tend to have a smaller physical footprint than B-mode transducers. Lower sound wave frequencies permit deeper tissue penetration, but at the sacrifice of higher resolution [121]. Higher frequency devices provide higher resolution of points at the same depth, but hold shallower tissue penetration capability compared to lower frequency transducers [121]. Generally 7.5 or 10 MHz has been the most popular choices employed in prior FSTT studies, however, there has been a range from 1 to 13MHz: see e.g., 1-4 MHz [16, 24]; 5 MHz [124]; 7.5 MHz [23, 122, 125]; 10MHz [19, 70, 87, 88]; 11.4 MHz [86]; 5-13 MHz [123].

There have been some attempts to use ultrasound with the face placed prone into a water bath to avoid transducer contact [126-128], but it is worth noting that this may deform the soft tissues (see e.g., Fig. 3) under the complex effects of gravity, holding one's breath, and the face's own buoyancy in water. Typically, 3D ultrasound images are acquired with this approach, from which a 2D median slice is extracted for FSTT measurement [126-128].

Brief History & Recent Advances: Ultrasound was first used in craniofacial identification by Russian [84] and German [16] teams, however, Helmer was the first to publish German FSTT tables associated with this work [16]. Initially, ultrasound machines were large and clunky (Fig. 8), which made smaller more portable A-mode devices the most popular initial choice for FSTT measurement [16, 24, 73]. Recent technological advances have decreased the size of the B-mode ultrasound machine to now only a transducer handpiece that plugs directly into a tablet, making B-mode devices much more user-friendly, more portable and more popular (Fig. 8).

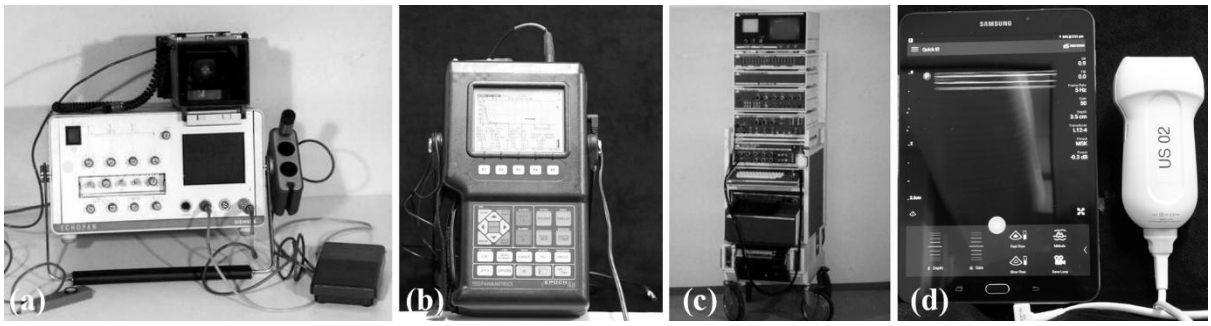

Fig. 8: Old and new ultrasound technology: (a) Siemens Echopan A-mode ultrasound device as used by Helmer [16] (imaged reproduced from the German Ultrasound Museum [http://www.ultraschallmuseum.de/] with permission from B. Frentzel-Beyme); and (b) Epoch 4b A-mode ultrasound flaw detector (Panametrics, Waltham, USA) as used by De Greef et al. [70, 73] (c) 1975 B-mode ATL Mark III device (imaged reproduced from the German Ultrasound Museum [http://www.ultraschallmuseum.de/] with permission from B. Frentzel-Beyme); and (d) modern-day (2018) hand held Philips Lumify B-mode device attached to a handheld tablet [Bothell, WA, USA] (L12-4).

Data fidelity: Living subjects are easily imaged in upright posture using ultrasound, which is a distinct advantage, but note here that a variety of body positions have been used in prior FSTT studies. For example, Helmer's hallmark study using A-mode ultrasound used subjects in a reclined seated position [16]; Hodson et al. measured participants in the supine position [122]; Smith et al.

have used prone positions [126-128]; and others have used seated, but fully upright subject orientations [23, 87, 88, 123].

As ultrasound methods do not involve ionising radiation, they are ideal for the measurement of children [23, 24, 86, 122] and repeated measurements of the same subjects in measurement error studies [73, 78, 87, 88]. One limitation is that ultrasound does not work well on embalmed tissues. As ultrasound images can be difficult to interpret, either by multiple reflectance peaks in A-mode, or image complexity in B-mode, operator expertise is an important consideration for data integrity [124]. Practitioner expertise is also important for proper transducer placement, such that the echo beam is orientated perpendicular to the tissue interfaces [87, 123] to avoid any overestimation of tissue depths [19]. It is important to note for imaging that, due to sound wave cancellation effects, the ultrasound beam emitted from the transducer is characterized by a near field converging beam (Fresnel zone) and a far field divergent beam (Fraunhofer zone) [121], i.e., the acoustic lens is not uniform along its length. There are a number of artefacts can affect the quality of ultrasound images including side lobes, reverberations, and aliasing [129]. Soft tissue compression is a major risk when measuring the delicate and easily moved soft tissues of the face, since ultrasound requires physical contact with the hand held transducer [16, 23, 73, 87, 126].

It is worth noting that the largest sampled single FSTT study so far conducted (n = 967 adults aged > 18 years) was undertaken by De Greef et al. [70] using A-mode ultrasound.

‘Plain film’ Radiographic Cephalograms:

Synopsis:

In this method, the subject stands side-on to an X-ray unit, facing directly forward and at a specific distance from the X-ray device. Following exposure to the X-ray source, a radiographic image is recorded on an X-ray sensitive film/detector. The less dense soft tissues down the midline of the face can be differentiated from the underlying denser skull due to differential X-ray absorption by the tissues (Fig. 9). Some variations of the method such as xeroradiography, which uses a specialised

image receptor plate, offer edge enhancement capabilities (Fig. 9). FSTT measurements can be taken on the X-ray pictures (lateral cephalograms) at specific anatomical landmarks and these measurements converted to life-size using magnification factors calculated from the subject-to-source distance [33].

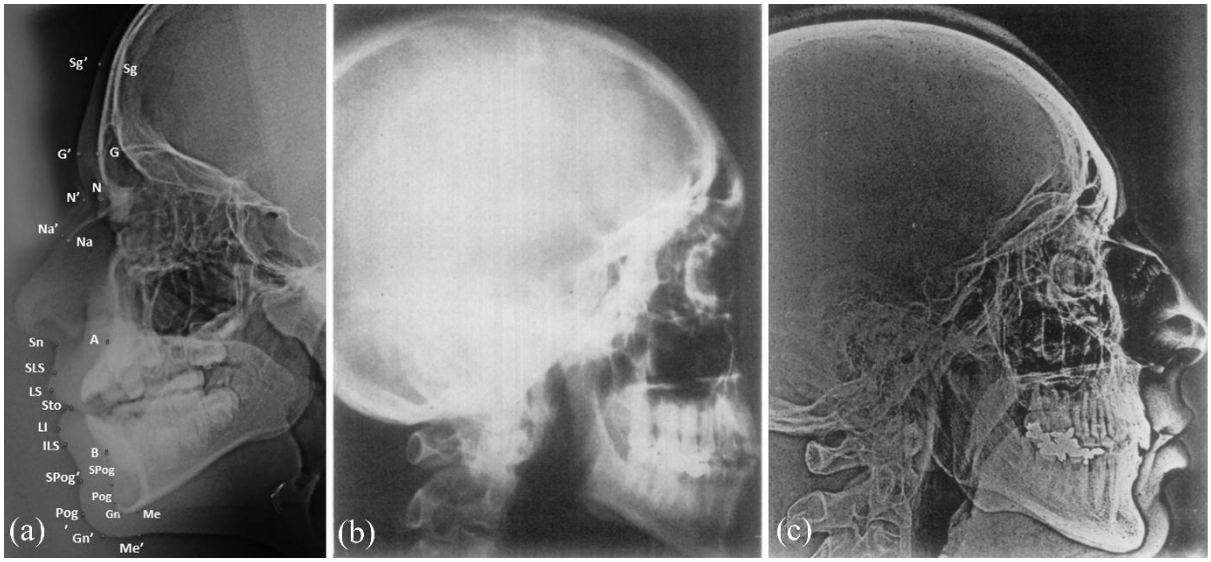

Fig. 9: Radiography: (a) modern-day diagnostic cephalogram with landmarks highlighted for FSTT measurement after Gibelli et al. [40]; imaged reproduced from [40] with permission by Elsevier; (b) film radiograph of a head taken using 1970's radiographic equipment; imaged reproduced from [130] with permission by Elsevier; and (c) comparable 1970's xeroradiographic image to demonstrate edge enhancement that accompanies xeroradiography making skull and skin surface easier to differentiate; imaged reproduced from [130] with permission by Elsevier.

Variations of Technique: Subjects are normally positioned in the Frankfurt Horizontal using a cephalostat [131]. Frontal view radiographs may sometimes be acquired either in addition to, or in place of, lateral views. A radiographic grid placed between the subject and the radiation source may be used to attenuate radiation scatter and, thereby, produce clearer radiographic images [120]. Radio-opaque barium [19], bismuth [132], or even small lead shot pellets [132] have been applied to the face ahead of imaging to make skin surface landmarks more readily visible on the radiograph.

As mentioned above, xeroradiography is also possible using reusable selenium plates [120, 133], though this method was phased-out in the late 1980's. This method was especially popular in dental specialities [133] (and mammography) and is notable for its edge enhancement [120], positive-mode image and print rather than film media [120].

FSTT measurements may be taken directly on the radiographic image, but it should also be noted that a common variation is to use tracings of the radiographs made on acetate paper over a light box, see e.g. [32, 33, 40]. This introduces yet another opportunity for error that is important to quantify in addition to the separate components of imaging and physical manipulation of measurement devices for data acquisition. As for other methods, including needle puncture, investigators may use different angles from the skull at which the FSTTs are taken [1, 87, 88, 96].

Brief History & Recent Developments: As for needle puncture, the origins of this method's use in craniofacial identification can be traced back to Welcker [134]. Welcker obtained facial soft tissue thicknesses at his own nasal bridge, using X-rays approximately one year after radiography was first developed by Röntgen [135].

As X-ray imaging technology has improved over the years, the radiation exposure times have decreased making the method safer for the participants being imaged. Image resolutions have also increased and recordings have moved from wet-film to direct digital recording formats, removing the requirement for wet-film processing. As lateral cephalograms are routinely taken for clinical dental assessments, much of the work using lateral radiographs has occurred in the orthodontics domain [29, 96, 136-138] holding ramifications again for the randomness of samples since the bulk of data is derived from clinical patients [20].

Data fidelity: Lateral cephalograms provided the first means of measuring living subjects in the upright position and free of any instrument contact with the skin surface—a significant advantage [96].

Subjects should be radiographed under standardized conditions (e.g., using a cephalostat and at standard distances), so that an appropriate correction factor for magnification can be applied – often around 10% [96]. One limitation lateral radiographs hold, is that locations lateral to the midline of the face are much harder to identify due to the superimposition of anatomical structures. While radio-opaque barium applied to the face ahead of the image capture can help facilitate these measurements [19], generally radiographic FSTT studies focus on midline landmarks only [1]. Since X-rays are generated from what is close to a single point source, perspective distortion in the image is another potential complication given relatively short source-to-subject distances that are employed [139-141].

Radiographs have been used to record FSTTs of cadavers [60], which suffers from the limitations of recording deceased participants not in a fully erect posture as described above for other methods. Measurement error assessments are difficult to obtain on living subjects when using radiographs because they require double exposure of subjects to the ionising radiation, which is normally avoided due to safety concerns [1]. This can be mitigated, to some degree, by combining diagnostic procedures with FSTT scans [22], where re-imaging may be part of clinical treatment requirements. Generally, these contexts are rare, meaning that error studies of radiographs rarely analyse the full measurement procedure. Instead, studies commonly ignore the imaging component and entirely focus on physical remeasurement of only one set of radiographic images—a major disadvantage—and they often concern non-random samples.

Computed Tomography (CT):

Synopsis:

CT produces cross-sectional images of the body with the aid of X-rays taken at multiple angles around the body. An X-ray tube and detectors are arranged on a gantry and opposite each other, with the patient positioned between the two, allowing the CT device (tube and detectors) to rotate around the object to be imaged. Typically, the patient occupies the supine position and can be moved incrementally along their longitudinal axis (z direction) to produce multiple axial slices via a

mechanically controlled table. Depending on the machine technology the patient may move through the gantry in a start-stop fashion or incrementally in a smooth continuous fashion as during helical scanning [121].

The raw orthoslices of conventional CT represent 2D greyscale axial sections (of a pre-specified thickness). The intensity of the X-rays registered at the detectors, at different positions around the subject, are used by the computer to mathematically construct (typically via back projection algorithms [129]) the cross-sectional images [142]. Modern CT images are typically of 512 x 512 pixels in size [129]. Multiple serial 2D images acquired as an image stack, can be reconstructed as 3D volume data by computer modelling software (Fig. 10).

The resulting base elements of the CT volume data are small 3D cubes called voxels [143]. Each voxel holds a numerical value expressed in Hounsfield Units (HU), just like the 2D pixel data in the 2D orthoslices when viewed axially. The Hounsfield scale refers to the attenuation of an x-ray beam on different materials normalized with respect to the coefficient of attenuation in water (0 HU) and air (-1000 HU) [121]. The Hounsfield scale has an interval from -1000 to +1000 [143], where each HU represents 0.1% of the attenuation of water [143]. In the human body, many tissues differ to each other in terms of their Hounsfield units allowing differentiation/segmentation between most (but not all) anatomical structures [143]. E.g., fat is approximately -100 HU, soft tissues are often between +20 to +70 HU and bone is generally > +400 HU [143].

Variations of Technique: No matter what CT scanner is used, the patient's CT experience tends to be similar—lying down on a movable CT table that slides into the gantry. Whilst there are differences in machine technology for image acquisition, such as 1^{st}, 2^{nd}, 3^{rd}, and 4^{th} generation machines, all variants ultimately produce images comprised of voxels as a result of the gantry and the subject being moved relative to one another to acquire serial images. Difference in shape of the X-ray beam, e.g., cone beam versus fan, is another variation (see section Cone Beam Computed Tomography).

The choice of CT machine for scanning is important because it determines both speed of the scan acquisition and the amount of radiation to which subjects are exposed [121]. Different investigators may also choose different CT settings at the time of scanning, for example the thickness and number of slices to be acquired, which in part sets different scan resolutions. Thinner slices require a larger serial slice number to cover the same field of view, they increase the scan time, they increase the radiation dose to the subject, but they also provide higher tissue imaging resolutions. A good example of variations in slice data in the FSTT context is provided by Parks et al. where 12 different protocols with slice thicknesses ranging from 0.98 mm to 6.0 mm are described within a single study [144].

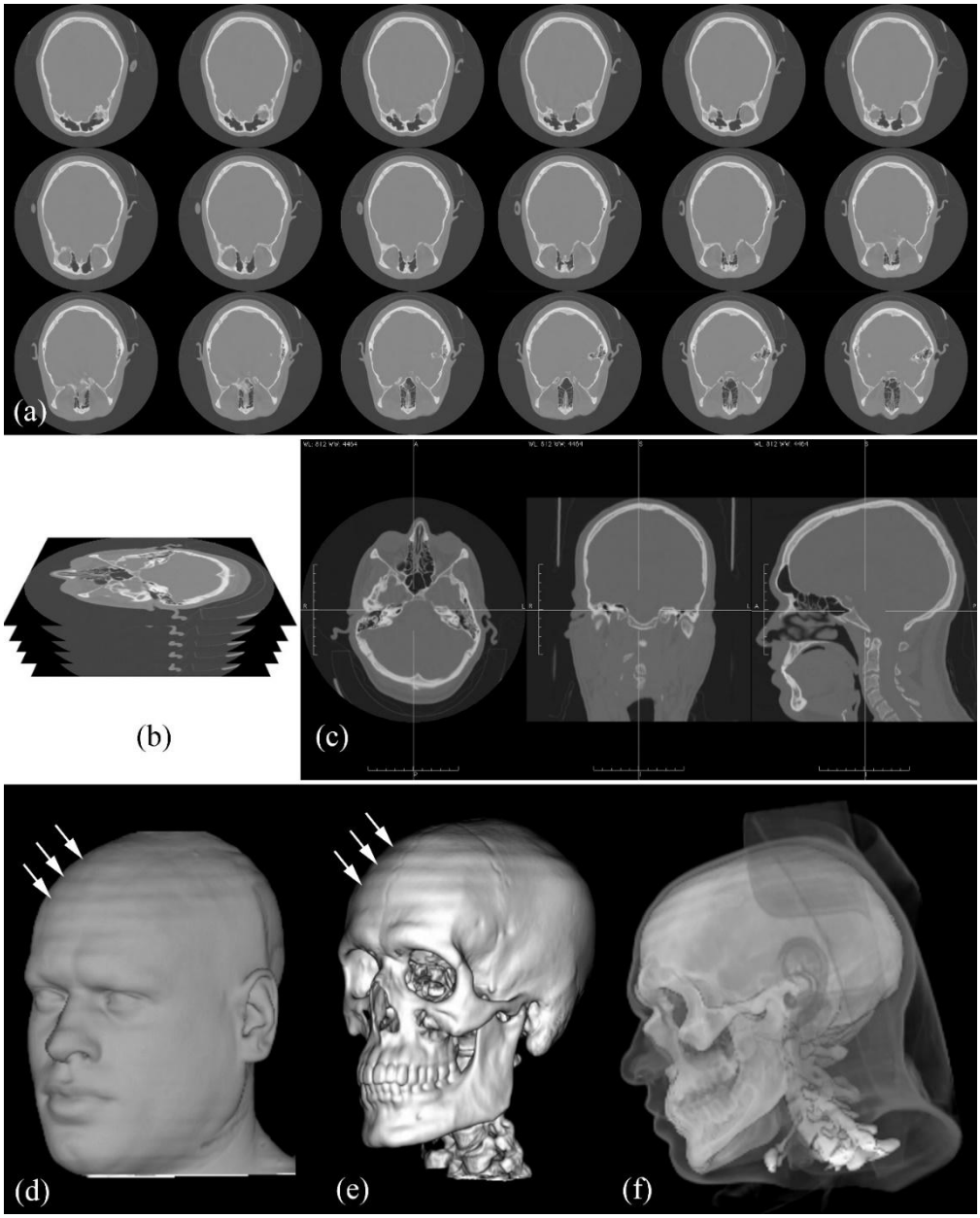

Fig. 10: Example CT images: (a) sequential axial orthoslices of a head CT taken with a spiral GM Light Speed CT scanner at 1.25 mm slice thickness and viewed in side-by-side fashion. Here only 18 images from a full head scan consisting of 32 serial images are displayed; (b) illustration of how the 2D orthoslice images represent a stack of images, captured in sequence along the z-axis of the body; (c) mathematical analysis of the image stacks enables 2D reconstructions to be generated in user-specified planes including oblique and transverse (left), coronal (middle), and parasagittal (right). These views are termed multi-planar reconstructions since sections are calculated from many orthoslices; (d) 3D volume render of the skin surface (white arrows mark stair step artefact due to helical CT); (e) 3D volume render of the skull alone (white arrows mark stair step artefact due to helical CT); and (f) 3D volume render of the skull with overlying skin surface at transparency. All images have been generated using OsiriX v.4.1.2.

It should be noted that the same body tissues will not generate the same CT values if scanned with different CT scanners, that is, the Hounsfield scale expression of the tissues differs between CT scanners and with varying energies on the same CT scanner [145]. This can complicate image segmentations since a single threshold value for segmentation of any single anatomical structure may produce a noisy result within and between devices.

Brief History & Recent Advances: The first CT scanners were used in the clinical setting in 1972 [121, 129] and used a single detector with a rotate-translate motion of the body to acquire further sequential axial slices [142]. The first whole-body CT scanner was produced in 1974 [129]. These early machines were superseded by multi-detector devices that enabled multiple image acquisition with single tube/detector rotations decreasing scan time [142]. The first helical CTs were developed in the late 1980s (also known as spiral or volume CT) [129] with the development of slip-ring technology that enabled constant electrical supply without fixed cables to the spinning detector [142]. Over the years this technology improved, again with multiple detectors being added, which by the late 1990s enabled larger volumes to be captured faster with higher resolution and reduced radiation dosages [142]. With regards to slice capabilities 4-, 16- and 32-slices were initially common but now

represent obsolete technology with 64- and 128-slice machines readily available. At the top end of the commercial market, 640-slice machines are available [146]. Dual source CT also exists to enable tissues with similar HU to be better differentiated [142].

In the craniofacial identification context, the first use of CT was by Phillips and Smuts in 1996 using a Elcint 2400 CT scanner [22]. FSTTs of both deceased and living individuals have been investigated with CT since 2007; for studies of deceased individuals see [147, 148], while for living individuals see [45, 47, 48, 50, 51, 54, 95, 144, 149-151].

Data fidelity: CT provides FSTT measurement of living individuals [1] and is a non-contact approach [1, 49], which are prime advantages. CT scans provide excellent morphological details and contrast between hard and soft tissue, which is why the segmentation of the skull is more easily accomplished using CT than MRI [95, 149]. The CT and 3D multislice reconstruction methods enable FSTTs to be taken at any point on the face [1], in sparse or dense manner [152]. These images, just like ultrasound or other radiographic recordings, enable measurements to be undertaken at any time [93].

Disadvantages of CT include radiation exposure [95], approximately 2 mSv or 0.2 effective radiation dose for head scans in adults [142], incomplete scans of the region of interest (ROI) [50], misalignment of hard tissue landmarks and their soft tissue equivalent due to the supine position of the patient [71, 72, 74, 153], distortion of the facial soft tissues due to a head strap placed over the forehead, or by the subject's face being compressed by oxygen masks in clinical scans [77] and limited scan resolution (e.g., large slice thickness), particularly on old technology [50, 153]. It is worth noting that picture quality may be degraded by artefacts such as aliasing, beam hardening, cupping and streaking—the later resulting from registration artefacts or patient movement [119, 121, 129]. Helical CT is also subject to stairstep artefacts (see Fig. 10d/e) due to equipment rotation during simultaneous Z-axis travel [129].

The accuracy of the CT data depends on scan resolution, both in the transverse plane (transaxial resolution) and longitudinally along the length of the subject (Z-axis) [121]. The transaxial or in-plane resolution is determined (amongst other things [129]) by the detector size or width—small detector size gives higher resolution [121]. The Z-sensitivity is determined by the slice width [121]. All else being equal, thinner slices yield higher resolution, but also higher noise [121]. Larger slices yield larger steps in the surfaces of the 3D reconstructions (see e.g., Fig. 11). It should be recognized that lower resolution CT scans often possess difficulties in acquiring and retaining thin cortical bone regions in skulls (orbits, maxillae, pterion etc.), such that these areas may be incomplete or absent on scans on some individuals to varying degrees (see e.g., Fig. 10 & 11).

CT images hold advantages to traditional 2D X-rays because the superimposition of body structures on one plane is avoided, instead organs can be visualized either as 3D reconstructions, 2D reconstructions, or as original 2D orthoslices [121]. In earlier work where multi-slice reconstruction was not possible, or was not undertaken, FSTT values should be considered with care since the orientation of the raw orthoslices used for measurement may not adhere to standard measurement planes [77], see e.g., [22, 45, 77, 149, 150].

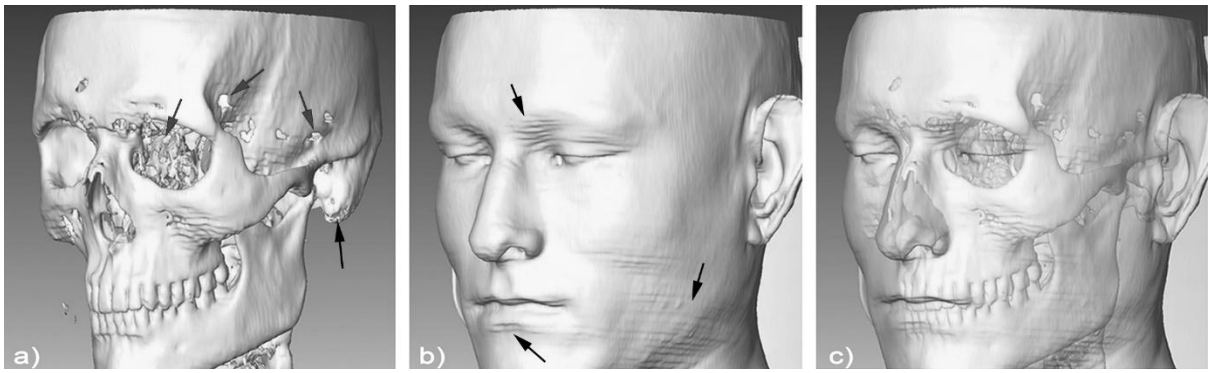

Fig. 11: 3D volume render of skull (a) and face (b) from CT scans recorded on a Phillips Mx8000 spiral CT scanner (Amsterdam, Netherlands) using a voxel resolution of 0.5mm. Image (c) shows visualisation of the skull through the face at opacity. Grey arrows in (a) indicate imaging artefacts such as holes in skull where bones are thin. Black arrows in (b) illustrate step artefacts (also present in

a)) that result from image slices and impact accurate representation of smooth skull and face surfaces. Images adapted from [51] with permission from Elsevier.

FSTT measurements have been taken between landmarks or at right angles to the bone surface or at angles bisecting the bony surface [114], while others have measured tissue depths in planes of particular head positions (such as when measurements are taken from raw 2D CT orthoslices without multi-slice reconstruction; see citations above). Guyomarc´h et al. [50] proposed an unconventional method of measuring FSTT parallel with the Frankfurt Horizontal plane in an attempt to provide improved standardisation; an approach which was also used by Thiemann et al. [54].

Cone Beam Computed Tomography (CBCT):

Synopsis:

A variation of traditional CT, CBCT uses a pyramidal or cone-shaped beam, which encompasses a large field of view (FOV) [121] that in turn is recorded on a large flat X-ray receptor [154]. In contrast, regular CT uses a fan or wedge shaped field of radiation [154] that is received on narrow, curved, linear array(s) of X-ray receptors [121]. As such, the scan is projected in three dimensions in CBCT, rather than multiple two-dimensional slices stacked atop one another as is usual practice in traditional CT [154].

Similar to traditional CT scans, CBCT relies on an x-ray emitter and detector that rotate around the subject [154], but unlike traditional CT methods, which require multiple rotations to be conducted at various levels, CBCT only requires one full or partial rotation (depending on the required FOV) [154]. Subject to machine type, the emitter in CBCT will either release a constant beam of radiation whilst rotation is occurring, or a sequence of radiation pulses [154]. A salient advantage of CBCT over traditional CT method, especially for FSTT data acquisition, is the ability to measure participants in an upright position, thus acquiring face morphology with a gravity vector directed inferiorly relative to the standard anatomical position.

Variation of Techniques: Depending on the CBCT machine that is used for the measurements, different parameters may be used by different operators. Primary among these is the voxel size used during the scan. Published voxel sizes used in FSTT investigations range typically from 0.3 to 0.4 mm [49, 93, 118]. Another important aspect to consider is the FOV size. Somewhat dependent on the capabilities of individual machines, the FOV can change drastically between studies, with Fourie et al. using a 17 mm FOV [118], versus a much larger 200 x 179 mm FOV used by Hwang et al. [93]. Of course, inherent with imaging modalities such as CT, radiographs and CBCT is the exposure time of the patient. Thus far, only Hwang et al. have reported an exposure time used to obtain scans specifically for FSTT measurement (= 17 seconds [93]).

Brief History & Recent Advances: CBCT is relatively new technology that has had limited use so far in the craniofacial identification domain—only four studies so far employ this technology [49, 93, 118, 155]—it has excellent future potential. The first use of this technology in craniofacial identification can be traced to Masoune et al. [155]. Because the technique is relatively new, there have been minimal advances since its first use. Advances mainly appear in the clarity of the scans and reduced radiation doses to the patients, moving from a continual beam of x-ray emission to a pulsed emission [154]. The 3D images produced by CBCT look very similar to those produced by regular CT multi-slice reconstruction (see Fig. 11).

Data fidelity: CBCT allows subjects to be measured in an upright position [49, 93], which is a significant advantage in comparison to other medical imaging technologies such as CT and MRI. As a non-contact method, CBCT ensures that there is no risk of soft tissue compression during FSTT measurement. Use of a conical or pyramidal-shaped beam of radiation increases the FOV of a single scan, thus negating the requirement for separate scans stacked atop one another as is practice in traditional CT images and potential for error during reconstruction [154]. Because of the reduced time required for acquisition of CBCT scans compared to traditional CT scans and decreased opportunity for involuntary patient movement during the scan, the CBCT images hold advantages of fewer motion artefacts [154]. The radiation dose of CBCT is significantly lower than the radiation dose attributed to

traditional CT scans, which is another prime advantage [93]. These benefits are counterbalanced by lower contrast resolution on CBCT scans in contrast to regular CT [156].

As clinical patients often form FSTT study cohorts with CBCT [22], this again encourages sample bias due to non-random sample selection. Challenges around accurate segmentation at tissue boundaries are retained in CBCT as for CT, which may impact on accuracy of FSTT measurements [93]. Like other radiography based approaches, surface skin tones are not retained in the images.

Magnetic Resonnance Imaging (MRI):

Synopsis:

MRI or nuclear spin tomography uses strong magnetic fields and radiowaves to generate images of the body [120]. Unlike CT scanning, it does not involve ionizing radiation, which is a marked advantage, but acquisition of MRI images tends to be more expensive. The magnetic field and radio waves are used to polarize hydrogen atoms in the body (make them spin the same way) then the protons are measured as they flip back to their equilibrium state at different times between different tissues to generate the image [120]. In MRI, the bone tissue appears very dark (black) and the soft tissues hold higher luminosity—the opposite of CT scans (compare Fig. 12 to Fig. 10).

Fig. 12: Parasagittal MRI reconstruction used for FSTT measurement by Sahni et al. [91]. Note black outline of skull. Image reproduced from [91] with permission from Elsevier.

Variations of Technique: Machines offering different resolutions may be used (see Brief History & Recent Advances). Measurements may also be taken on the raw MRI slices or on multi-slice reconstructed images as per CT methods described above.

Brief History & Recent Advances: MRI was developed in the mid-1970s [157]. The first investigator to use MRI for FSTT measurement was Helmer [158]. MRI was further utilized for FSTT measurement by Sahni et al. [90,91] and more recently by Sandamini et al. [114].

Advances in machinery include increases in strength of magnetic field able to be generated, as measured in Tesla (T), e.g., 7T as opposed to 3T systems [143, 161-163]. The newer 7T systems provide higher signal-to-noise ratios than 3T systems [161, 162] and improve image resolutions from approximately 1 mm to 0.5 mm [163]. 7T systems were first implemented in clinical settings in the early 2000's and have been especially popular for brain imaging in neuroscience [161-163].

Data fidelity: As for needle puncture methods and CT, MRI is limited because most MRI machines are designed to scan subjects in the supine position [91]. Therefore, MRI data can be expected to possess all the usual complications associated with supine subject scans—larger volumes around the eyes and cheeks and smaller volumes along the nasolabial fold [71] (Fig. 4).

MRI holds the advantage that it is non-contact and soft tissues can be very well visualized [143]. As for CT, multi-slice reconstructed MRI images tend to have lower resolutions than raw slice images. Measurement from raw slices may be problematic if the head is not correctly orientated to standard planes [77]. Segmentation and differentiation of the bone from the soft tissue is broadly recognized to be much easier and better in CT images than MRI, at least using present day algorithms [74]. For recent advances regarding bone correction factors, such as zero echo time (ZTE) in PET/MR see [164]. Per other imaging methods, MRI is also subject to image artefacts including: susceptibility, motion, coil, machine-dependent and gradient field [121].

FSTT Measurement Approaches that Combine Imaging Methods

Synopsis:

Rather than using one imaging modality alone to take FSTT measurements, two imaging modalities have been jointly used to separately image the skull and the facial surface before combining the two to take FSTT [165]. This holds the benefit that the best modalities for each tissue component can be utilised (e.g., soft tissues acquired in upright positions, which is not relevant to skull tissue imaging), but it adds a subsequent requirement to use a good registration function, so that errors are not introduced when aligning the scans [165].

Brief History, Variations of Technique & Recent Advances: This approach was first undertaken in 2009 using low-dose CT and holographic facial images [165]. Since its first use, it has also subsequently been employed using traditional 16-slice spiral CT and 3D photographs taken using a Breuickmann FaceScanIII-180 device [76, 166].

Data fidelity: As mentioned above, benefit of these methods is that separate imaging modalities well-suited to skull and face capture can be employed. However, this necessitates precise registration methods and introduces complications that separate imaging sessions (at different times) may be used rather than simultaneous scan acquisition at the same session [167]. Prieels et al. [165, 167] recommend using the forehead and nasal bridge for image registration using iterative closest point, but no error quantification of misalignment is provided. Kustar et al. [76], also using iterative closest point registration, recommended the nose and forehead region for alignment with eight registration landmarks obtaining a mean error of 0.37 mm, $s = 0.33$ for 40 subjects (whether or not this mean represents calculations from signed or unsigned data is not clear despite multiple requests to the publishing authors).

FSTT Measurement Errors

One might expect the FSTT literature to be awash in data on measurement errors given the vast array of different measurement variations employed as detailed above—but this is not the case. Of all 90 FSTT papers published (either in English or easily accessible to us) from 1883 to 2018, less than half (42%) make any effort to address measurement error (Tables 1-4). This increases marginally to 56% of 36 papers published in the last 5 years. Of those papers attempting to document measurement error, the vast majority (95%) do not report error statistics that cover the full breadth of data collection protocol [29-31, 33-40, 42, 45, 47-54, 91, 93]. Generally, the studies concerned here are those that employ ionizing radiation methods (12 of 13 lateral radiograph studies & 12 of 12 CT/CBCT studies). Subsequently, in these cases the reported measurement error is very likely underestimated. For lateral cephalograms, the mean relative technical error of measurement (rTEM) for *remeasurement of the same (not repeat) images* ranges between 1.1 to 8.6 % [34]. For CT, the rTEM for intra-observer error for the remeasurement of single images varies between 0.15-1.92% [54]; and for inter-observer error is between 0.17-3.91 % [54]. Unfortunately, until these methods include repeated image acquisition the full extent of errors involved in these measurement methods will remain unknown. Values derived from single images in the range of 4-8% can only indicate that the overarching measurement errors, which will be even larger, are not negligible.

<Table 2 about here>

Only 5% of studies use two separate data acquisition sessions to assess the full measurement protocol and utilise the technical error of measurement (TEM) for error calculation [168]. These studies pertain only to needle puncture [65, 67] and B-mode ultrasound [87, 88]. For needle puncture, the mean error magnitudes (rTEM) for intra-observer error varies between 10-15 % (maximum at any landmark was 34% or in millimetre terms 3.4 mm [65]); and for inter-observer error is 12% (maximum at any landmark was 34 % or in millimetre terms 6.4 mm [65]). For B-mode ultrasound, the mean error

magnitudes (rTEM) for intra-observer error vary between 8-10 % (maximum at any landmark was 15 % or in millimetre terms 1.7 mm [88]); and for inter-observer error vary between 11-21 % (maximum at any landmark was 45 % or in millimetre terms 5.6 mm [88]). These measurement error metrics, which assess the entire data acquisition process are, expectedly, substantially higher than error reports for ionizing radiation methods where only one half of the measurement protocol is assessed. While repeat scans for MRI are possible (scans are radiation free), no such repeat studies examining the error embedded in measurement procedures have yet been conducted—perhaps because cost is prohibitive.

<Table 3 about here>

It is important to note that in some cases, entirely inappropriate statistical methods have been used to assess the measurement error, e.g., one way ANOVA for independent (not dependant) samples [53] and a surprisingly large number of cephalogram studies attempt to report measurement errors using t-tests [29, 32, 33, 39, 40, 49, 73, 91, 122]. Student's t-tests should be avoided when examining measurement errors since these tests will only identify systematic error, not normally distributed errors about a mean of zero [168], which most of the measurement error is likely to be. Not surprisingly then, studies employing t-tests (including Hodson's first use for FSTT measurement errors in 1985) rarely find statistically significant differences between repeated measurement sessions [29, 32, 33, 39, 40, 49, 73, 91, 122]. Counter to what prior studies claim, the lack of statistical significance following t-testing is not a reliable indicator of little or negligible error because the test is insensitive to error normally distributed about a mean of zero [168].

<Table 4 about here>

Interclass correlation coefficients (ICC) have also been popularly used for describing the measurement error. While valid, care needs to be taken with this statistic since large ICCs, e.g., well above 0.9 (see Table 1-4), though positive, do not mean that the error is negligible [168]. ICC values can be high when substantial measurement error is present [168]. Statements that 'perfect reliability' exists based on high (0.98), but not perfect ICC values, is misleading, especially when repeated imaging sessions are not used for the analysis, see e.g., [42].

<Table 5 about here>

In summary, measurement errors for full data collection protocols have only been conducted in a limited number of studies applicable to two of the eight main measurement methods (needle puncture and ultrasound). These data indicate that measurement errors are in the vicinity of 10 % across both modalities of measurement and can be as high as 45 %. Even for medical imaging methods, where repeat scans are not acquired (i.e., only single images remeasured), errors are as high as 2-8%. Measurement errors exceeding 1 mm clearly invalidate any temptation to report FSTT means to multiple decimal places as popular in the current research literature [19] and they also strongly question near universal subdivision of samples based on small millimetre or submillimetre differences [1, 20, 169].

Recommendations

1. Any new FSTT study should report its corresponding measurement error by taking repeated measurements of the same subjects using the full data collection protocol. Where ionizing radiation is used (precluding repeat scanning of healthy subjects), then alternate permissible methods of rescanning should be investigated e.g., using cadaver surrogates, artificial surrogates, or clinical samples where rescanning subjects with time delay between scans is required as part of treatment. Measurement errors should be reported for commonly measured landmarks (e.g., as revealed by meta-analysis [1, 2, 170]) to facilitate comparisons at the same landmarks between different measurement modalities and studies. If the measurement error cannot be established using the full data collection protocol, then this must be clearly stated in the written work.
2. The measurement error should be, at least, described by the TEM (or rTEM) because these statistics flag the error well [168, 171-174]. If other statistics are employed such as the raw mean difference, the mean absolute error, r and ICC, then they should be provided *in addition* to the TEM and rTEM. There is no harm in reporting the full suite of error metrics to be comprehensive. Student's t-tests should not be used for evaluating the measurement error because the error is commonly distributed about a mean of zero such that the t-tests will produce statistically insignificant results even though substantial measurement error exists [168]. Where mean errors are reported, it should be specified if they are calculated from signed or unsigned data, as different signed errors will tend to average out when in opposing directions. Care should be taken not to report data to unreasonable degrees of decimal places (i.e., two or three), when the measurement error clearly does not justify this degree of precision [1, 19, 20].
3. When FSTT means are calculated from the raw data and presented as point estimators for new subjects, they should be accompanied with their associated estimation errors [4]. The standard error of the estimate (SEE) should be used as this statistic and will in part include measurement errors embedded in the training data used to produce the estimation model. The

estimation error can be calculated by testing the means as point estimates for known subjects that have not contributed to the estimation model [4].

New FSTT studies reporting means as point estimators should not be published unless accompanied by estimation error metrics that document their accuracy (such as the SEE). Ideally, any newly published means should offer estimation improvements over and above pre-existing means. To help facilitate validation tests, we provide free and open source R tool (*TDValidator)* at CRANIOFACIALidentification.com. This tool enables investigators to enter their FSTT means, for standard T-Table landmarks, and test their accuracy of estimation using previously acquired FSTT data for other known individuals in the C-Table data repository. *TDValidator* reports three error metrics: raw mean error (mm and %), mean absolute error (mm and in %) and the SEE.

References

[1] C.N. Stephan and E.K. Simpson, Facial soft tissue depths in craniofacial identification (part I): an analytical review of the published adult data. J. Forensic Sci. 53 (2008) 1257-1272.

[2] C.N. Stephan, 2018 Tallied facial soft tissue thicknesses: adult and sub-adult data. Forensic Sci. Int. 280 (2017) 113-123.

[3] H. Welcker, Schiller's Schädel und Todtenmaske, nebst Mittheilungen über Schädel und Todtenmaske Kant's, Viehweg and Son, Braunschweig, 1883.

[4] C.N. Stephan, Accuracies of facial soft tissue depth means for estimating ground truth skin surfaces in forensic craniofacial identification. Int. J. Legal Med. 129 (2015) 877-888.

[5] C.N. Stephan, TDStats - a capability for standardized facial soft tissue thickness analysis in R. Forensic Sci. Int. 289 (2018) 304-309.

[6] K.T. Taylor, Forensic Art and Illustration, CRC Press, Boca Raton, FL, 2001.

[7] B.P. Gatliff, Facial sculpture on the skull for identification. Am. J. Forensic Med. Path. 5 (1984) 327-332.

[8] B.P. Gatliff and C.C. Snow, From skull to visage. The Journal of Biocommunication 6 (1979) 27-30.

[9] B.P. Gatliff and K.T. Taylor, Three-dimensional facial reconstruction on the skull, in: K.T. Taylor (Ed.), Forensic Art and Illustration, CRC Press, Boca Raton, FL, 2001, pp. 419-475.

[10] M.M. Gerasimov, Osnovy Vostanovleniia Litsa po Cherepo, Izdat. Akademii Nauk SSSR., Moskva, 1949.

[11] M.M. Gerasimov, Vosstanovlenie lica po cerepu, Izdat. Akademii Nauk SSSR, Moskva, 1955.

[12] J. Prag and R. Neave, Making Faces: Using Forensic and Archaeological Evidence, British Museum Press, London, U.K., 1997.

[13] C.N. Stephan and M. Henneberg, Building faces from dry skulls: Are they recognized above chance rates? J. Forensic Sci. 46 (2001) 432-440.

[14] C. Wilkinson, Forensic Facial Reconstruction, Cambridge University Press, Cambridge, U.K., 2004.

[15] R.G. Taylor and C. Angel, Facial reconstruction and approximation, in: J.G. Clement and D.L. Ranson (Eds.), Craniofacial Identification in Forensic Medicine, Oxford University Press, New York, 1998, pp. 177-185.

[16] R. Helmer, Schädelidentifizierung durch elekroniesche Bildmischung: Zugleich ein Beitrag zur Konstitutionsbiometrie und Dickenmessung der Gesichtsweichteile., Krminalistik-Verlag, Heidelberg, 1984.

[17] C.N. Stephan, R.G. Taylor and J.A. Taylor, Methods of facial approximation and skull-face superimposition, with special consideration of method development in Australia, in: M. Oxenham (Ed.), Forensic Approaches to Death, Disaster and Abuse, Australian Academic Press, Bowen Hills, 2008, pp. 133-154.

[18] A.M. Brues, Identification of skeletal remains. J. Crim. Law Criminol. Pol. Sci. 48 (1958) 551-556.

[19] W.A. Aulsebrook, P.J. Becker and M.Y. İşcan, Facial soft-tissue thickness in the adult male Zulu. Forensic Sci. Int. 79 (1996) 83-102.

[20] C.N. Stephan, L. Munn and J. Caple, Facial soft tissue thicknesses: noise, signal and P. Forensic Sci. Int. 257 (2015) 114-122.

[21] C.N. Stephan and E.K. Simpson, Facial soft tissue depths in craniofacial identification (part II): an analytical review of the published sub-adult data. J. Forensic Sci. 53 (2008) 1273-1279.

[22] V.M. Phillips and N.A. Smuts, Facial reconstruction: Utilization of computerized tomography to measure facial tissue thickness in a mixed racial population. Forensic Sci. Int. 83 (1996) 51-59.

[23] M.H. Manhein, G.A. Listi, R.E. Barsley, R. Musselman, N.E. Barrow and D.H. Ubelaker, In vivo facial tissue depth measurements for children and adults. J. Forensic Sci. 45 (2000) 48-60.

[24] C.M. Wilkinson, In vivo facial tissue depth measurements for White British children. J. Forensic Sci. 47 (2002) 459-465.

[25] S. De Greef, D. Vandermeulen, P. Claes, P. Suetens and G. Willems, The influence of sex, age and body mass index on facial soft tissue depths. Forensic Sci. Med. Path. 5 (2009) 60-65.

[26] S. De Greef and G. Willems, Three-dimensional cranio-facial reconstruction in forensic identification: latest progress and new tendencies in the 21st century. J. Forensic Sci. 50 (2005) 12-17.

[27] W.-J. Lee, S. Mackenzie and C.M. Wilkinson, Facial identification of the dead, in: S. Black and E. Ferguson (Eds.), Forensic Anthropology: 2000-2010, CRC Press: Taylor and Francis Group, Boca Raton, 2011, pp. 363-394.

[28] C.N. Stephan, E.K. Simpson and J.E. Byrd, Facial soft tissue depth statistics and enhanced point estimators for craniofacial identification: the debut of the shorth and the 75-shormax. J. Forensic Sci. 58 (2013) 1439-1457.

[29] E.R. Dumont, Mid-facial tissue depths of white children: an aid in facial feature reconstruction. J. Forensic Sci. 31 (1986) 1463-1469.

[30] J.S. Genecov, P.M. Sinclair and P.C. Dechow, Development of the nose and soft tissue profile. Angle Orthod. 60 (1990) 191-198.

[31] K. Kasai, Soft tissue adaptability to hard tissues in facial profile. Am. J. Dentofacial Orthoped. 113 (1998) 674-684.

[32] T.N. Garlie and S.R. Saunders, Midline facial tissue thicknesses of subadults from a longitudinal radiographic study. J. Forensic Sci. 44 (1999) 61-67.

[33] M.A. Williamson, S.P. Nawrocki and T.A. Rathbun, Variation in midfacial tissue thickness of African-American children. J. Forensic Sci. 47 (2002) 25-31.

[34] A. Kurkcuoglu, C. Pelin, B. Ozener, R. Zagyapan, Z. Sahinoglu and A.C. Yazici, Facial soft tissue thickness in individuals with different occlusion patterns in adult Turkish subjects. Homo 62 (2011) 288-297.

[35] T.M.F. Fernandes, A. Pinzan, R. Sathler, M. Roberto de Freitas, G. Janson and F.P. Vieira, Comparative study of the soft tissue of young Japanese-Brazilian, Caucasian and Mongoloid patients. Dental Press J. Orthod. 18 (2013) 116-124.

[36] N. Briers, T.M. Briers, P.J. Becker and M. Steyn, Soft tissue thickness values for Black and Coloured South African children aged 6 to 13 years. Forensic Sci. Int. 252 (2015) 188.e181-188.e110.

[37] W. Jeelani, M. Fida and A. Shaikh, Facial soft tissue thickness among various vertical facial patterns in adult Pakistani subjects. Forensic Sci. Int. 257 (2015) 517.e511-517.e516.

[38] J. Wang, X. Zhao, C. Mi and I. Raza, The study on facial soft tissue thickness using Han population in Xinjiang. Forensic Sci. Int. 266 (2016) 585.e581-585.e585.

[39] V.S. Kotrashetti and M.D. Mallapur, Radiographic assessment of facial soft tissue thickness in South Indian population - An anthropologic study. J. Forensic Legal Med. 39 (2016) 161-168.

[40] D. Gibelli, F. Collini, D. Porta, M. Zago, C. Dolci, C. Cattaneo and C. Sforza, Variations of midfacial soft-tissue thickness in subjects aged between 6 and 18 years for the reconstruction of the profile: A study on an Italian sample. Legal Med. 22 (2016) 68-74.

[41] W. Jeelani, M. Fida and A. Shaikh, Age and sex-related variations in facial soft tissue thickness in a sample of Pakistani children. Aust. J. Forensic Sci. 49 (2017) 45-58.

[42] S.K. Buyuk, E. Genc, H. Simsek and A. Karaman, Analysis of facial soft tissue values and cranial skeletal widths in different body mass index percentile adolescent subjects. CRANIO (2018)

[43] F. Ayoub, M. Saadeh, G. Rouhana and R. Haddad, Midsagittal facial soft tissue thickness norms in an adult mediterranean population. Forensic Sci. Int. (2018) Early online.

[44] K.-D. Kim, A. Ruprecht, G. Wang, J.B. Lee, D.V. Dawson and M.W. Vannier, Accuracy of facial soft tissue thickness measurements in personal computer-based multiplanar reconstructed computed tomographic images. Forensic Sci. Int. 155 (2005) 28-34.

[45] D. Cavanagh and M. Steyn, Facial reconstruction: Soft tissue thickness values for South African black females. Forensic Sci. Int. 206 (2011) 215.e211-215.e217.

[46] H.-S. Hwang, S.-Y. Choe, J.-S. Hwang, D.-N. Moon, Y. Hou, W.-J. Lee and C. Wilkinson, Reproducibility of Facial Soft Tissue Thickness Measurements Using Cone-Beam CT Images According to the Measurement Methods. J. Forensic Sci. 60 (2015) 957-965.

[47] P. Paneková, R. Beňuš, S. Masnicová, Z. Obertová and J. Grunt, Facial soft tissue thicknesses of the mid-face for Slovak population. Forensic Sci. Int. 220 (2012) 293.e291-293.e296.

[48] Y. Dong, L. Huang, Z. Feng, S. Bai, G. Wu and Y. Zhao, Influence of sex and body mass index on facial soft tissue thickness measurements of the northern Chinese adult population. Forensic Sci. Int. 222 (2012) 396.e391-396.e397.

[49] N.A. Perlaza Ruiz, Facial soft tissue thickness of Colombian adults. Forensic Sci. Int. 229 (2013) 160.e161-160.e169.

[50] P. Guyomarc'h, F. Santos, B. Dutailly and H. Coqueugniot, Facial soft tissue depths in French adults: variability, specificity and estimation. Forensic Sci. Int. 231 (2013) 411.e411-411.e410.

[51] O. Bulut, S. Sipahioglu and B. Hekimoglu, Facial soft tissue thickness database for craniofacial reconstruction in the Turkish adult population. Forensic Sci. Int. 242 (2014) 44-61.

[52] A. Drgáčová, J. Dupej and J. Velemínská, Facial soft tissue thicknesses in the present Czech population. Forensic Sci. Int. 260 (2016) 106.e101-106.e117.

[53] A. Lodha, M. Mehta, M.N. Patel and S.K. Menon, Facial soft tissue thickness database of Gujarati population for forensic craniofacial reconstruction. Egyptian J. Forensic Sci. 6 (2016) 126-134.

[54] N. Thiemann, V. Keil and U. Roy, In vivo facial soft tissue depths of a modern adult population from Germany. Int. J. Legal Med. 131 (2017) 1455-1488.

[55] D. Toneva, S. Nikolova, I. Georgiev, S. Harizanov, D. Zlatareva, V. Hadjidekov and N. Lazarov, Facial soft tissue thicknesses in Bulgarian adults: relation to sex, body mass index and bilateral asymmetry. Folia Morphol. 77 (2018) 570-582.

[56] W. His, Anatomische Forschungen über Johann Sebastian Bach's Gebeine und Antlitz nebst Bemerkungen über dessen Bilder. Abh. MathPhysikal. CI. Kgl. Sächs. Ges. Wiss. 22 (1895) 379-420.

[57] J. Kollmann and W. Büchly, Die Persistenz der Rassen und die Reconstruction der Physiognomie prähistorischer Schädel. Arch. Anthropol. 25 (1898) 329-359.

[58] J. Czekanowski, Untersuchungen über das Verhältnis der Kopfmaße zu den Schädelmaßen. Arch. Anthropol. 6 (1907) 42-89.

[59] F. Stadtmüller, Zur Beurteilung der plastischen Rekonstruktionsmethode der Physiognomie auf dem Schädel. Z. Morph. Anthropol. 22 (1922) 337-372.

[60] D. Leopold. Identifikation durch Schädeluntersuchung unter besonderer Berücksichtigung der Superprojektion. Leipzig: Karl-Marx-Universität, (1968).

[61] P.R.N. Sutton, Bizygomatic diameter: The thickness of the soft tissues over the zygions. Am. J. Phys. Anthropol. 30 (1969) 303-310.

[62] J.S. Rhine and C.E. Moore, Tables of facial tissue thickness of American Caucasoids in forensic anthropology. Maxwell Museum Technical Series 1 (1984)

[63] A.S. Forrest. An investigation into the relationship between facial soft tissue thickness and age in Australian Caucasion cadavers. Department of Anatomy. Brisbane: The University of Queensland, (1985).

[64] T. Blythe. A re-assessment of the Rhine and Moore Technique in Forensic Facial Reconstruction. Anatomical Sciences. Manchester: The University of Manchester, (1996).

[65] E. Simpson and M. Henneberg, Variation in soft-tissue thicknesses on the human face and their relation to craniometric dimensions. Am. J. Phys. Anthropol. 118 (2002) 121-133.

[66] M. Sutisno. Human facial soft-tissue thickness and its value in forensic facial reconstruction. Department of Pathology, Faculty of Medicine. Sydney: The University of Sydney, (2003).

[67] M. Domaracki and C.N. Stephan, Facial soft tissue thicknesses in Australian adult cadavers. J. Forensic Sci. 51 (2006) 5-10.

[68] J.F. O'Grady, R.G. Taylor and J.G. Clement. Facial tissue thickness: a study of cadavers in Melbourne. International Association of Forensic Science Scientific Symposium. Adelaide, (1990).

[69] J.G. Clement and D.L. Ranson, Craniofacial Identification in Forensic Medicine, Arnold, London, 1998.

[70] S. De Greef, P. Claes, D. Vandermeulen, W. Mollemans, P. Suetens and G. Willems, Large-scale in-vivo Caucasian soft tissue thickness database for craniofacial reconstruction. Forensic Sci. Int. 159S (2006) S126-S146.

[71] L. Munn and C.N. Stephan, Changes in face topography from supine-to-upright position—And soft tissue correction values for craniofacial identification. Forensic Sci. Int. 289 (2018) 40-50.

[72] U. Ozsoy, R. Sekerci and E. Ogut, Effect of sitting, standing, and supine body positions on facial soft tissue: detailed 3D analysis. Int. J. Oral Maxillofac. Surg. 44 (2015) 1309-1316.

[73] S. De Greef, P. Claes, W. Mollemans, M. Loubele, D. Vandermeulen, P. Suetens and G. Willems, Semi-automated ultrasound facial soft tissue depth registration: method and validation. J. Forensic Sci. 50 (2005) 1282-1288.

[74] D. Vandermeulen, P. Claes, S. De Greef, G. Willems, J. Clement and P. Suetens, Automated facial reconstruction, in: C.M. Wilkinson and C. Rynn (Eds.), Craniofacial Identification, Cambridge University Press, Cambridge, 2012, pp. 203-221.

[75] O. Bulut, C.-Y.J. Liu, F. Koca and C. Wilkinson, Comparison of three-dimensional facial morphology between upright and supine positions employing three-dimensional scanner from live subjects. Legal Med. 27 (2017) 32-37.

[76] A. Kustar, L. Forro, I. Kalina, F. Fazekas, S. Honti, S. Makra and M. Friess, FACE-R: A 3D database of 400 living individuals' full heat CT- and Face Scans and Preliminary GMM analysis for craniofacial reconstruction. J. Forensic Sci. 58 (2013) 1420-1428.

[77] J. Caple, C. Stephan, L. Gregory and D. MacGregor, Effect of head position on facial soft tissue depth measurements obtained using computed tomography. J. Forensic Sci. 61 (2016) 147-152.

[78] C.N. Stephan, R. Priesler, O. Bulut and M.B. Bennett, Turning the tables of sex distinction in craniofacial identification: why females possess thicker facial soft tissues than males, not vice versa. Am. J. Phys. Anthropol. 161 (2016) 283-295.

[79] C.N. Stephan, B. Meikle and M.B. Bennett, Variation in human facial soft tissue thicknesses by sex: females hold proportionately larger depths than males. (2019) In preparation.

[80] S. Codinha, Facial soft tissue thicknesses for the Portuguese adult population. Forensic Sci. Int. 184 (2009) 80.e81-80.e87.

[81] S. Rhine. Tissue thickness for Southwestern Indians. Maxwell Museum, Physical Anthropology Laboratories: University of New Mexico, (1983).

[82] J.S. Rhine and H.R. Campbell, Thickness of facial tissues in American blacks. J. Forensic Sci. 25 (1980) 847-858.

[83] S.V. Tedeschi-Oliveira, R.F.H. Melani, N. de Almeida and L.A. de Paiva, Facial soft tissue thickness of Braziallian adults. Forensic Sci. Int. 193 (2009) 127.e121-127.

[84] G.V. Lebedinskaya, T.S. Balueva and E.V. Veselovskaya, Principles of facial reconstruction, in: M.Y. İşcan and R.P. Helmer (Eds.), Forensic Analysis of the Skull, Wiley-Liss, New York, 1993, pp. 183-198.

[85] W.N. Chan, G.A. Listi and M.H. Manhein, In vivo facial tissue depth study of Chinese-American adults in New York City. J. Forensic Sci. 56 (2011) 350-358.

[86] L. Jia, B. Qi, J. Yang, W. Zhang, Y. Lu and H.-L. Zhang, Ultrasonic measurement of facial tissue depth in a Northern Chinese Han population. Forensic Sci. Int. 259 (2016) 247.e241-247.e246.

[87] C.N. Stephan and R. Preisler, In vivo facial soft tissue thicknesses of adult Australians. Forensic Sci. Int. 282 (2018) 220.e221-220.e212.

[88] C.N. Stephan and E. Sievwright, Facial soft tissue thickness (FSTT) estimation models—and the strength of correlations between craniometric dimensions and FSTTs. Forensic Sci. Int. 286 (2018) 128-140.

[89] S. Niinimaki and A. Karttunen. Finnish facial tissue thickness study. in: V.-P. Herva, (ed). Proceedings of the 22nd Nordic Archaeological Conference. University of Oulu: Gummerus Kirjapaino Oy, (2006) 343-352.

[90] D. Sahni, I. Jit, M. Gupta, P. Singh and S. Suri, Preliminary study on facial soft tissue thickness by magnetic resonance imaging in Northwest Indians. Forensic Science Communications 4 (2002)

[91] D. Sahni, Sanjeev, D. Singh, I. Jit and P. Singh, Facial soft tissue thickness in northwest Indian adults. Forensic Sci. Int. 176 (2008) 137-146.

[92] S. Sipahioglu, H. Ulubay and H.B. Diren, Midline facial soft tissue thickness database of Turkish population: MRI study. Forensic Sci. Int. 219 (2012) 282.e281-282.e238.

[93] H.-S. Hwang, M.-K. Park, W.-J. Lee, J.-H. Cho, B.-K. Kim and C.M. Wilkinson, Facial soft tissue thickness database for craniofacial reconstruction in Korean adults. J. Forensic Sci. 57 (2012) 1442-1447.

[94] J.H. Chung, H.T. Hsu, H.T. Chen, G.S. Huang and K.P. Shaw, A CT-scan database for the facial soft tissue thickness of Taiwan adults. Forensic Sci. Int. 253 (2015) 132.e131-132.e111.

[95] F. Tilotta, F. Richard, J. Glaunes, M. Berar, S. Gey, S. Verdeille, Y. Rozenholc and J.F. Gaudy, Construction and analysis of a head CT-scan database for craniofacial reconstruction. Forensic Sci. Int. 191 (2009) 112.e111-112.

[96] R.M. George, The lateral craniographic method of facial reconstruction. J. Forensic Sci. 32 (1987) 1305-1330.

[97] H. Helwin, Die Profilanalyse, eine Möglichkeit der Identifizierung unbekannter Schädel. Gegenbaurs Morphol. Jahrb. 113 (1969) 467-499.

[98] W. Weining. Röntgenologische Untersuchungen zur Bestimmung der Weichteildickenmaße des Gesichts. Zentrum der Rechtsmedizin des Klinikums. Frankfurt: Johann Wolfgang Goethe-Universität (1958).

[99] W. Anderson. The correlation between soft tissue thickness and bony proportions of the skull and how they relate to facial reconstruction. Department of Anatomical Sciences. Adelaide: The University of Adelaide, (1996).

[100] F. Birkner, Beiträge zur Rassenanatomie der Gesichtsweichteile. Corr. Bl. Anthrop. Ges. Jhg. 34 (1904) 163-165.

[101] A.N. Burkitt and G.H.S. Lightoller, Preliminary observations on the nose of the Australian aboriginal with a table of aboriginal head measurements. J. Anat. 57 (1923) 295-312.

[102] H.v. Eggeling, Anatomische Untersuchungen an den Köpfen von ver Hereros, einem Herero- und einem Hottentottenkind, in: L. Schultze (Ed.), Forschungsreise im westlichen und zentralen Südafrika, Denkschriften, Jena, 1909, pp. 323-348.

[103] E. Fischer, Anatomische Untersuchungen an den Kopfweichteilen zweier Papua. Corr. BL. Anthrop. Ges. Jhg. 36 (1905) 118-122.

[104] I.M. Bankovski. Die Bedeutung der Unterkieferform und-stellung für die photographische Schädelidentifizierung. Frankfurt: Johann Wolfgang Goethe-Universität, (1958).

[105] H. Edelman, Die Profilanalyse: Eine Studie an photographischen und röntgenographischen Durchdringungsbildern. Z. Morph. Anthropol. 37 (1938) 166-188.

[106] H. Ogawa, Anatomical study on the Japanese head by X-ray cephalometry. J. Tokyo Dent. Col. Soc. [Shika Gakuho] 60 (1960) 17-34.

[107] W. His, Johann Sebastian Bach. Forschungen über dessen Grabstätte, Gebeine und Antlitz. Bericht an den Rath der Stadt Leipzig, FCW Vogel, Leipzig, 1895.

[108] D. Berger. Untersuchungen über die Weichteildickenmaße des Gesichts. Institut für gerichtliche und soziale Medizin. Frankfurt/Main, Germany: Johann Wolfgang Goethe-Universität, 1965.

[109] G.I.C. Suazo, L.M. Cantín, M.D.A. Zavando, R.F.J. Perez and M.S.R. Torres, Comparisons in soft-tissue thicknesses on the human face in fresh and embalmed corpses using needle puncture method. Int. J. Morphol. 26 (2008) 165-169.

[110] N.H. de Almeida, E. Michel-Crosato, L.A. de Paiva and M.G. Biazevic, Facial soft tissue thickness in the Brazilian population: new reference data and anatomical landmarks. Forensic Sci. Int. 231 (2013) 404.e401-407.

[111] I. Robetti, M. Iorio and V. Mascaro, Die Stärke des Weichgewebes des Gesichtes zur Personenidentifizierung. Z. Rechtsmed. 89 (1982) 119-124.

[112] V.J. DiMaio and D. DiMaio, Forensic Pathology, CRC Press, Boca Raton, 2001.

[113] J. Caple and C.N. Stephan, A standardized nomenclature for craniofacial and facial anthropometry. Int. J. Legal Med. 130 (2016) 863-879.

[114] H. Sandamini, A. Jayawardena, L. Batuwitage, R. Rajapakse, D. Karunaratne, M. Vidanapathirana and A. Pallewatte, Facial soft tissue thickness trends for selected age groups of Sri Lankan adult population. Forensic Sci. Int. (2018) Early online.

[115] H. Suzuki, On the thickness of the soft parts of the Japanese face. J. Anthropol. Soc. Nippon 60 (1948) 7-11.

[116] T.D. Stewart, Essentials of forensic anthropology: Especially as developed in the United States, Charles C Thomas, Sprinfield, IL, 1979.

[117] V. Suk, Fallacies of anthropological identifications. Publications de la Facultae des sciences de l'Universitae Masaryk 207 (1935) 3-18.

[118] Z. Fourie, J. Damstra, P.O. Gerrits and Y. Ren, Accuracy and reliability of facial soft tissue depth measurements using cone beam computer tomography. Forensic Sci. Int. 199 (2010) 9-14.

[119] P.N.T. Wells, Scientific Basis of Medical Imaging, Churchill Livingstone, Edinburgh, 1982.

[120] J. Ball and T. Price. Chesneys' Radiographic Imaging. Oxford: Blackwell Science, 1995.

[121] J.T. Bushberg, J.A. Seibert, E.M. Leidholdt Jr and J.M. Boone, The Essential Physics of Medical Imaging, Lippincott, Williams & Wilkins, Philadelphia, 2012.

[122] G. Hodson, L.S. Lieberman and P. Wright, In *vivo* measurements of facial tissue thicknesses in American caucasoid children. J. Forensic Sci. 30 (1985) 1100-1112.

[123] L.J. Baillie, S. Ali Mirijali, B.E. Niven, P. Blyth and G.J. Dias, Ancestry and BMI influences on facial soft tissue depths for a cohort of Chinese and Caucasoid women in Dunedin, New Zealand. J. Forensic Sci. 60 (2015) 1146-1154.

[124] L.J. Baillie, J.C. Muirhead, P. Blyth, B.E. Niven and G.J. Dias, Position effect on facial soft tissue depths: a sonographic investigation. J. Forensic Sci. 61 (2015) S60-S70.

[125] I.H. El-Mehallawi and E.M. Soliman, Ultrasonic assessment of facial soft tissue thickness in adult Egyptians. Forensic Sci. Int. 117 (2001) 99-107.

[126] S.L. Smith and G.S. Throckmorton, A new technique for three-dimensional ultrasound scanning of facial tissues. J. Forensic Sci. 49 (2004) 1-7.

[127] S.L. Smith, G.S. Throckmorton and P.H. Buschang, A new method for measuring soft tissue thicknesses of the face using ultrasound. Am. J. Phys. Anthropol. 123 (2004) 184-185.

[128] S.L. Smith and G.S. Throckmorton, Comparability of radiographic and 3D-ultrasound measurements of facial midline tissue depths. J. Forensic Sci. 51 (2006) 244-247.

[129] P. Suetens, Fundamentals of Medical Imaging, Cambridge University Press, Cambridge, 2017.

[130] W.H. Binnie, A.J. Stacey, R. Davis and R.A. Cawson, Applications of xeroradiography in dentistry. J. Dent. 3 (1975) 99-104.

[131] B.S. Phulari, An Atlas on Cephalometric Landmarks, Jaypee Brothers Medical Publishers, New Delhi, 2013.

[132] J. Köstler. Röntgenstereoskopische Messungen der Weichteildicken in der Medianebene des Gesichtes an zwanzig jungen Personen weiblichen Geschlechtes: Friedrich Alexanders Universität Erlangen, 1940.

[133] B.J. Michelow and B. Guyuron, The chin: skeletal and soft-tissue components. Plast. Reconst. Surg. 95 (1995) 473-478.

[134] H. Welcker, Das Profil des menschlichen Schädels mit Röntgenstrahlen am Lebenden dargestellt. Korrespondenz-Blatt der Deutschen Gesellschaft für Anthropologie Ethnologie und Urgeschichte 27 (1896) 38-39.

[135] W.C. Röntgen, On a new kind of rays. Nature 53 (1896) 274-277. An Engish translation of an article that appeared in the Sitzungsberichte der Würzburger physikalisch-medizinische Gesellschaft, 1895.

[136] K.-V. Sarnäs and B. Solow, Early adult changes in the skeletal and soft-tissue profile. Eur. J. Orthod. 2 (1980) 1-12.

[137] R.S. Nanda, H. Meng, S. Kapila and J. Goorhuis, Growth changes in the soft tissue facial profile. Angle Orthod. 60 (1990) 177-190.

[138] W.A. Formby, R.S. Nanda and G.F. Currier, Longitudinal changes in the adult facial profile. Am. J. Orthod. Dentofacial Orthop. 105 (1994) 464-476.

[139] Z. Titlbach, Beitrage zur Bewertung der Superprojektionsmethode zur Identifizierung unbekannter Skelettfunde, in: Kriminalistik und forensische Wissenschaften, German Publisher of Sciences, Berlin, 1970, pp. 179-190.

[140] C. Stephan, Perspective distortion in craniofacial superimposition: logarithmic decay curves mapped mathematically and by practical experiment. Forensic Sci. Int. 257 (2015) e1-e8.

[141] C. Stephan and P. Guyomarc'h, Quantification of perspective-induced shape change of clavicles at radiography and 3D scanning to assist human identification. J. Forensic Sci. 59 (2014) 447-453.

[142] D.T. Ginat and R. Gupta, Advances in Computed Tomography Imaging Technology. Ann. Rev. Biomed. Eng. 16 (2014) 431-453.

[143] J.A. Weir, P.H., Imaging Atlas of Human Anatomy, Mosby, Edinburgh, 2003.

[144] C.L. Parks, A.H. Richard and K.L. Monson, Preliminary assessment of facial soft tissue thickness utilizing three-dimensional computed tomography models of living individuals. Forensic Sci. Int. 237 (2014) 146.e141-146.e110.

[145] A.A.T. Bui and R.K. Tara, Medical Imaging Informatics, Springer Science and Business, New York, 2010.

[146] S. Anim-Sampong, W.K. Antwi, B. Ohene-Botwe and R.S. Boateng, Comparison of 640-slice Aquilon ONE CT scanner's measured dosimetric parameters with ICRP dose reference levels for head, chest and abdominal CT examinations. Safety in Health 2 (2016) 1-8.

[147] D. Bellmann, T. Fuchs, A. Weidenbusch, J. Haber, K.M. Stein, T. Georg and J. Wilske, Computer-aided measurement of the tissue thickness of deceased persons with computer tomography scans of the head, in: T.M. Buzug, K.M. Sigl, J. Bongartz and K. Prüfer (Eds.), Facial Reconstruction: Forensic, Medical and Archaeological Methods of the Reconstruction of the Soft Facial Parts, Wolters Kluwer, Munich, 2007, pp. 21-39.

[148] R. Shimofusa, S. Yamamoto, T. Horikoshi, H. Yokota and H. Iwase, Applicability of facial soft tissue thickness measurements in 3-dimensionally reconstructed multi-detector-row CT images for forensic anthropological examination. Legal Med. 11 (2009) S256-S259.

[149] A.C. Frigo, O. Procopio, R. Peretta, G. Scattolin and G. Ferronato, Imaging of facial soft tissues in multislice computerized tomography: a new geometric method of analysis and its statistical validation. Oral. Surg. Oral. Med. Oral. Pathol. Oral. Radiol. Endodont. 110 (2010) 101-109.

[150] T. Saxena, S.R. Panat, N.C. Sangamesh, A. Choudhary, A. Aggarwal and N. Yadav, Facial soft tissue thickness in North Indian adult population. J. Indian Acad. Oral Med. Radiol. 24 (2012) 121-125.

[151] K.-S. Cha, Soft-tissue thickness of South Korean adults with normal facial profiles. Korean J.f Orthod. 43 (2013) 178-185.

[152] W. Shui, M. Zhou, Q. Deng, Z. Wu, Y. Ji, K. Li, T. He and H. Jiang, Densely calculated facial soft tissue thickness for craniofacial reconstruction in Chinese adults. Forensic Sci. Int. 266 (2016) 573.e571-573.e512.

[153] P. Claes, D. Vandermeulen, S. De Greef, G. Willems, J.G. Clement and P. Suetens, Computerized craniofacial reconstruction: Conceptual framework and review. Forensic Sci. Int. 201 (2010) 138-145.

[154] W.C. Scarfe and A.G. Farman, What is cone-beam CT and how does it work? Dent. Clinics North Am. 52 (2008) 707-730.

[155] J. Masoume, E. Farzad and M.M. Mohaddeseh, Facial soft tissue thickness in North-West of Iran. Adv. Biosci. Clin. Med. 3 (2015) 29-34.

[156] R. Gupta, A.C. Cheung, S.H. Bartling, J. Lisauskas, M. Grasruck, C. Leidecker, B. Schmidt, T. Flohr and T.J. Brady, Flat-panel volume CT: fundamental principles, technology, and applications. Radiographics 28 (2008) 2009–2022.

[157] P. Mansfield and P.K. Grannell, "Diffraction" and microscopy in solids and liquids by NMR. Phys. Rev. B 12 (1975)

[158] R. Helmer, F. Koschorek, B. Terwey and T. Frauen, Dickenmessung der Gesichtsweichteile mit Hilfe der Kernspin-Tomographie zum Zwecke der Identifizierung. Arch. Kriminol. 178 (1986) 139-150.

[159] F. Chen, Y. Chen, Y. Yu, Y. Qiang, M. Liu and D. Fulton, Age and sex related measurement of craniofacial soft tissue thickness and nasal profile in the Chinese population. Forensic Sci. Int. 212 (2011) 272.e271-272.e276.

[160] J. Vander Pluym, W.W. Shan, Z. Taher, C. Beaulieu, C. Plewes, A.E. Peterson, O.B. Beatties and J.S. Bamforth, Use of magnetic resonance imaging to measure facial soft tissue depth. Cleft Palate Craniofac. J. 44 (2007) 52-57.

[161] E. Springer, B. Dymerska, P.L. Cardoso, S.D. Robinson, C. Weisstanner, R. Wiest, B. Schmitt and S. Trattnig, Comparison of routine brain imaging at 3 T and 7 T. Invest. Radiol. 51 (2016) 469-482.

[162] H.C. Moon, H.-M. Baek and Y.S. Park, Comparison of 3 and 7 Tesla magnetic resonance imaging of obstructive hydrocephalus caused by tectal glioma. Brain Tumor Res. Treat. 4 (2016) 150-154.

[163] A. Nowogrodzki, The world's strongest MRI machines are pushing human imaging to new limits. Nature 563 (2018) 24-26.

[164] M. Khalifé, B. Fernandez, O. Jaubert, M. Soussan, V. Brulon, I. Buvat and C. Comtat, Subject-specific bone attenuation correction for brain PET/MR: can ZTE-MRI substitute CT scan accurately? Phys. Med. Biol. 62 (2017) 7814-7832.

[165] F. Prieels, S. Hirsch and P. Hering, Holographic topometry for a dense visualization of soft tissue for facial reconstruction. Forensic Sci. Med. Path. 5 (2009) 11-16.

[166] A. Kustar, Z. Gerendas, I. Kalina, F. Fazekas, B. Vari, S. Honti and S. Makra, FACE-R: 3D skull and face database for virtual anthropololgy research. Annales Historico-Naturales Musei Nationalis Hungarici 105 (2013) 313-319.

[167] F. Prieels. Holographic topometry with high resolution for forensic facial reconstruction. Medizinische Fakultät. Düsseldorf: Heinrich-Heine-Universität, (2009).

[168] H.S.M. Fancourt and C.N. Stephan, Error measurement in craniometrics: The comparative performance of four popular assessment methods using 2000 simulated cranial length datasets (g-op). Forensic Sci. Int. 285 (2018) 162-171.

[169] C.N. Stephan, R.M. Norris and M. Henneberg, Does sexual dimorphism in facial soft tissue depths justify sex distinction in craniofacial identification? J. Forensic Sci. 50 (2005) 513-518.

[170] C. Stephan, The application of the central limit theorem and the law of large numbers to facial soft tissue depths: T-Table robustness and trends since 2008. J. Forensic Sci. 59 (2014) 454-462.

[171] G. Dahlberg, Statistical methods for medical and biological students, George Allen & Unwin, London, 1940.

[172] T.A. Perini, G.L. de Oliveira, J.S. Ornellas and F.P. de Oliveira, Technical error of measurement in anthropometry. Rev. Bras. Med. Esporte 11 (2005) 86-90.

[173] T.R. Knapp, Technical error of measurement: a methodological critique. Am. J. Phys. Anthropol. 87 (1992) 235-236.

[174] S.J. Ulijaszek and D.A. Kerr, Anthropometric measurement error and the assessment of nutritional status. Brit. J. Nut. 82 (1999) 165-177.

Figure 1

Click here to download high resolution image

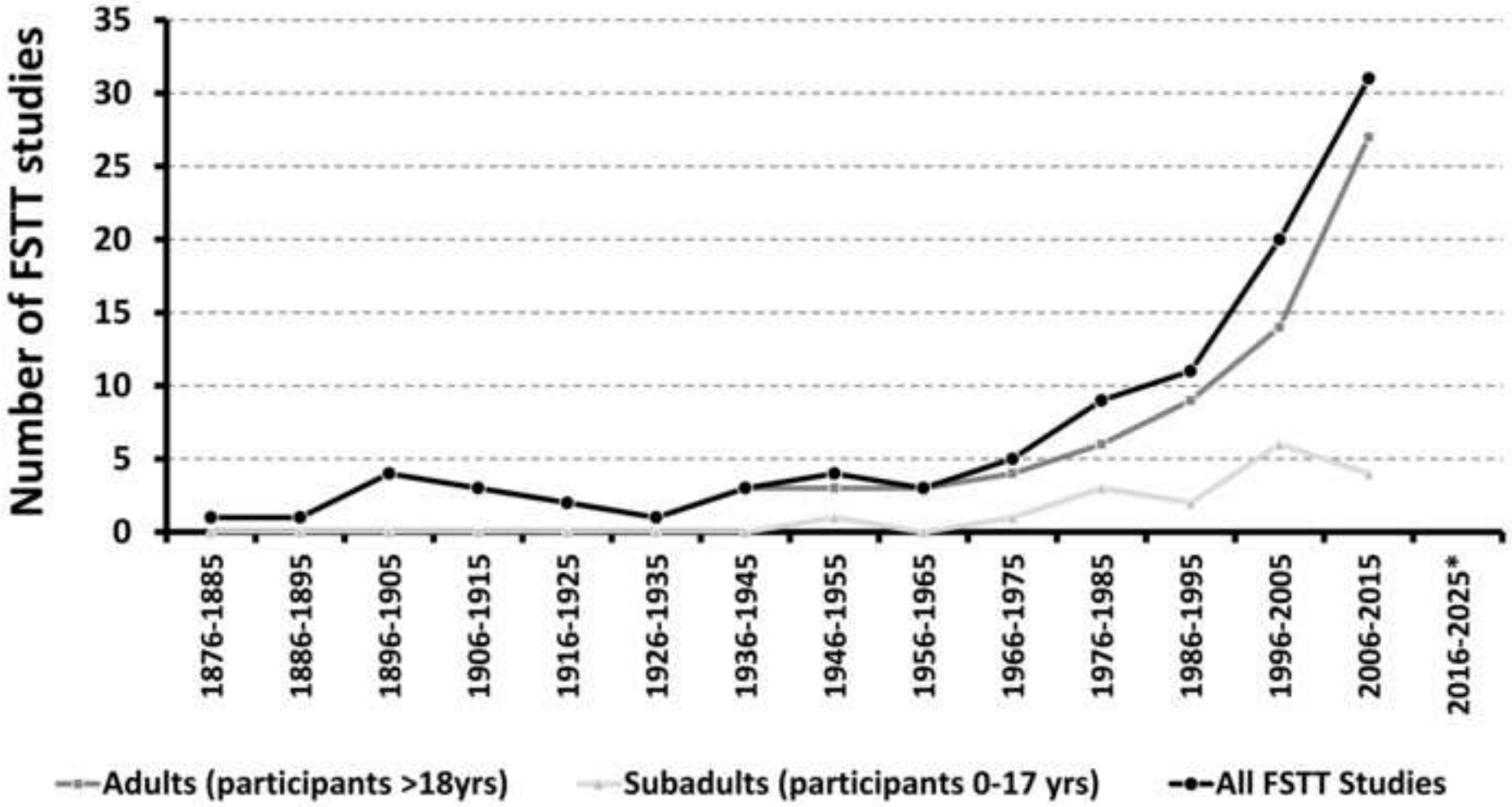

Figure 2

Click here to download high resolution image

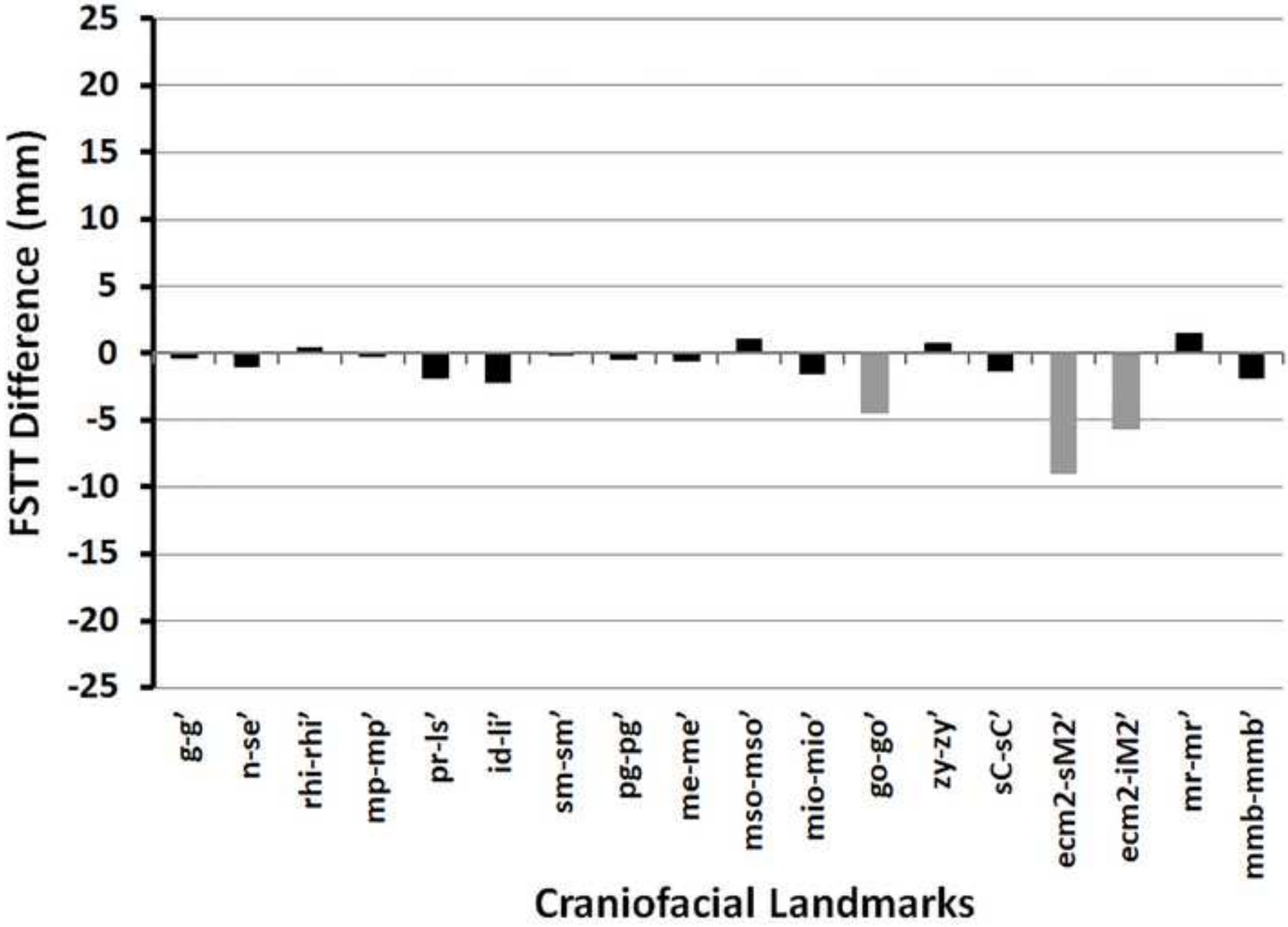

Figure 3

Click here to download high resolution image

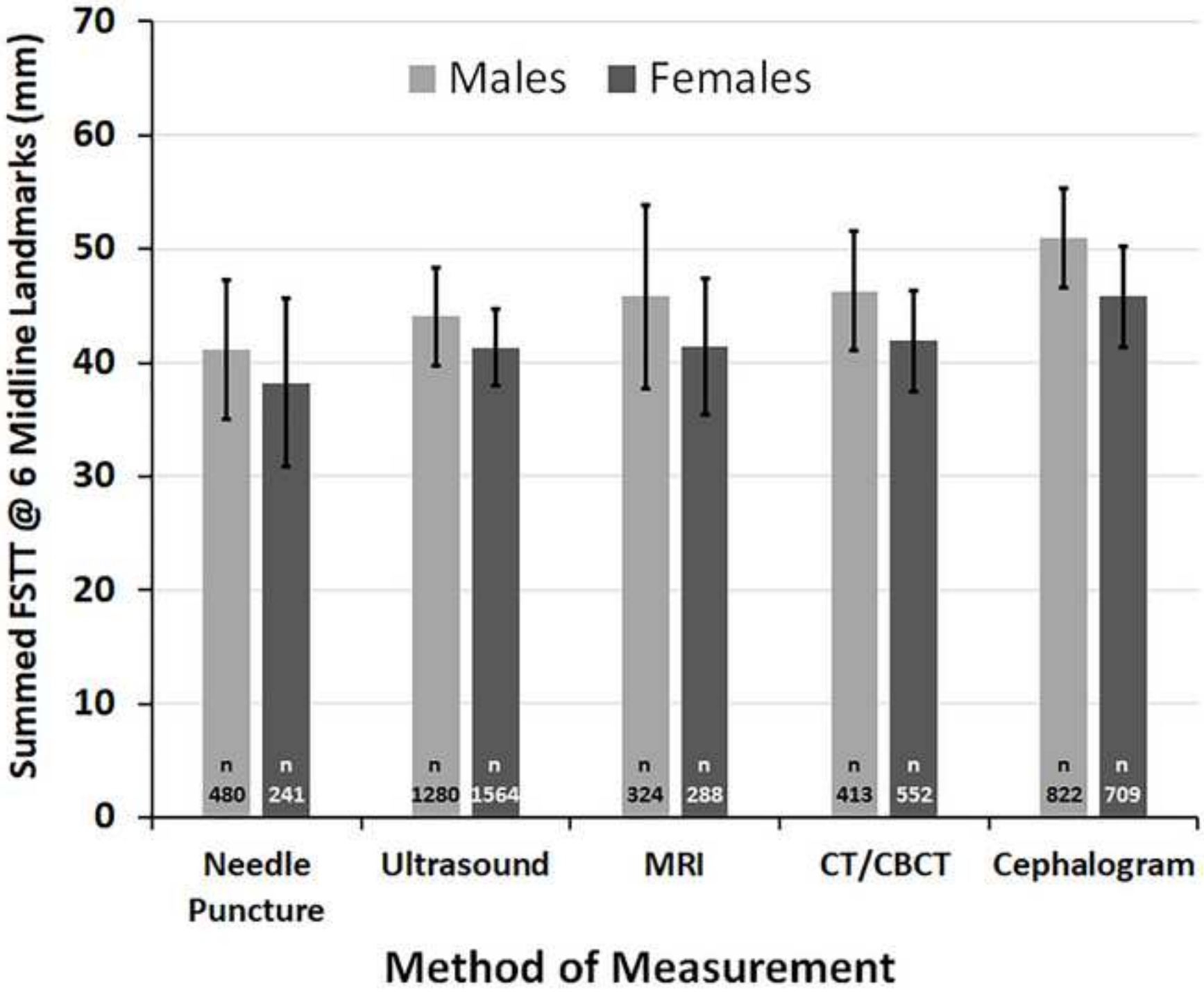

Figure 4

Click here to download high resolution image

Figure 5

Click here to download high resolution image

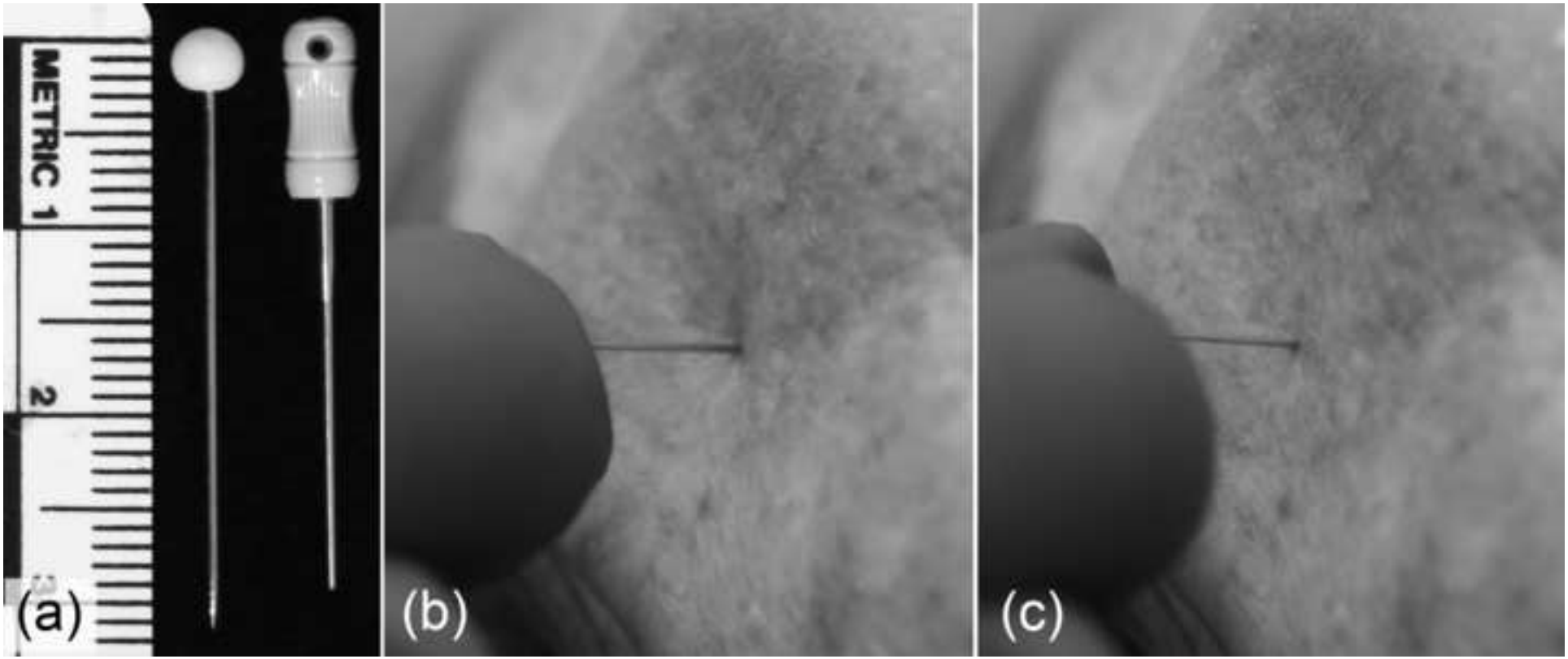

Figure 6

Click here to download high resolution image

Figure 7

Click here to download high resolution image

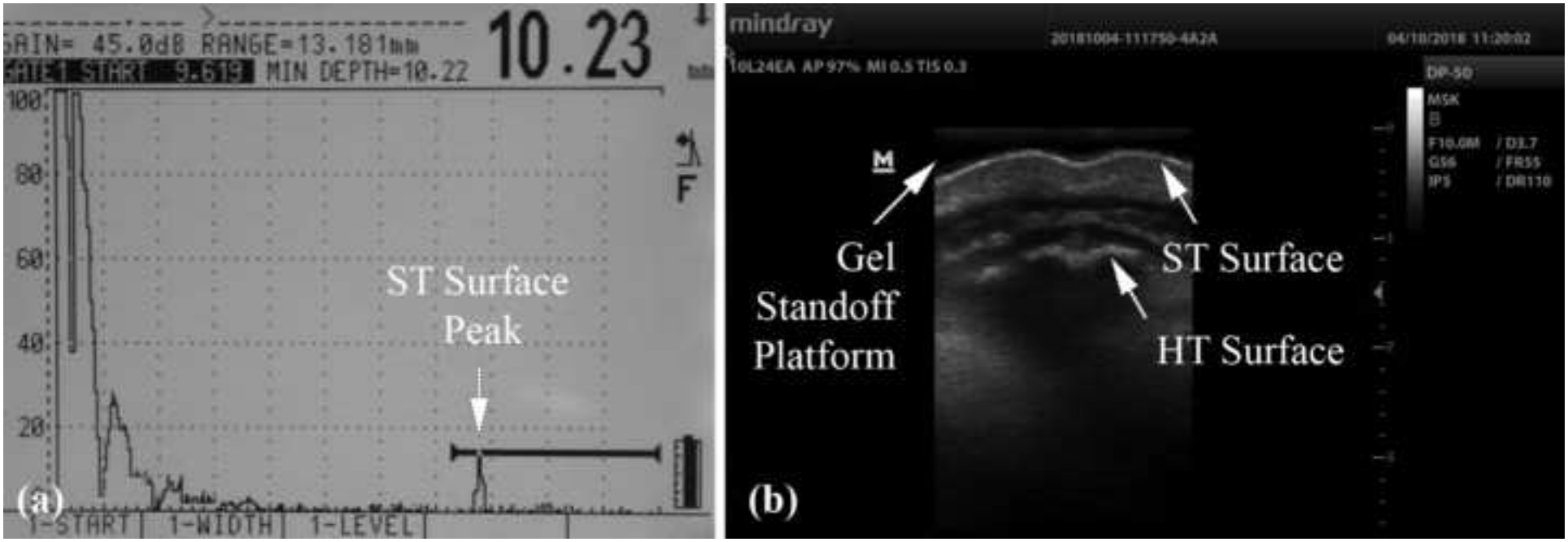

Figure 8

Click here to download high resolution image

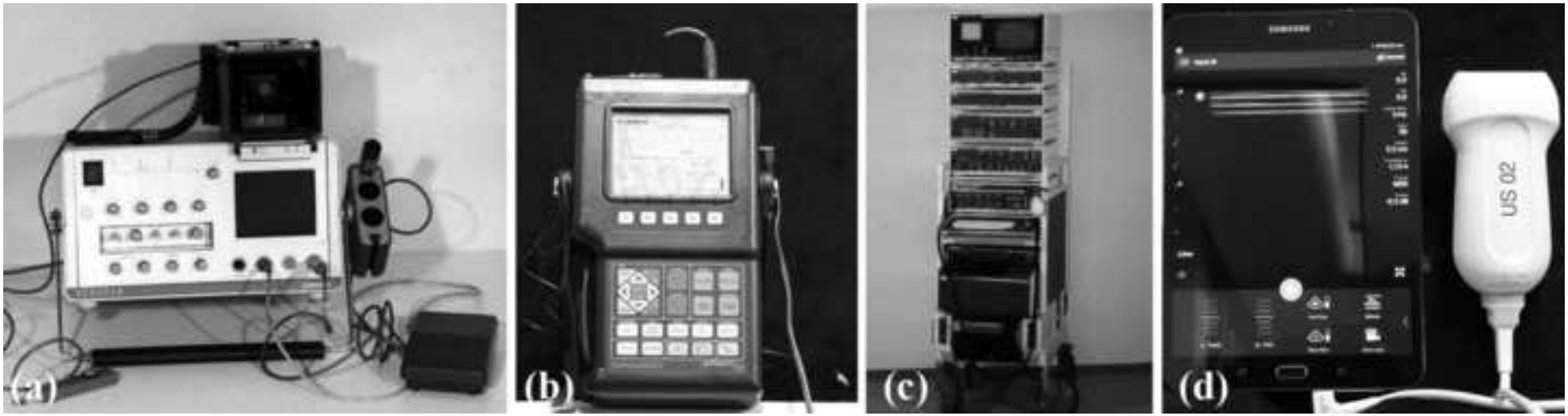

Figure 9

Click here to download high resolution image

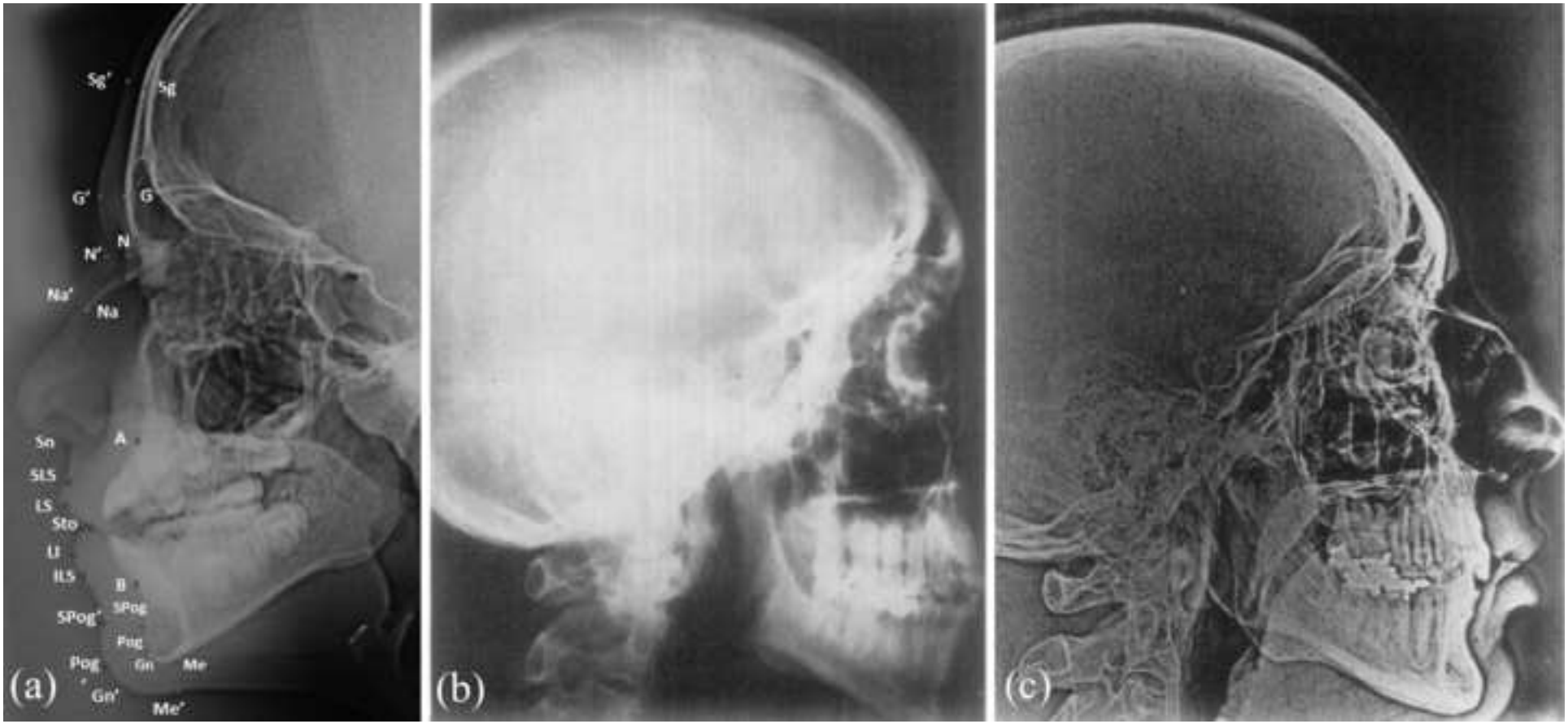

Figure 10

Click here to download high resolution image

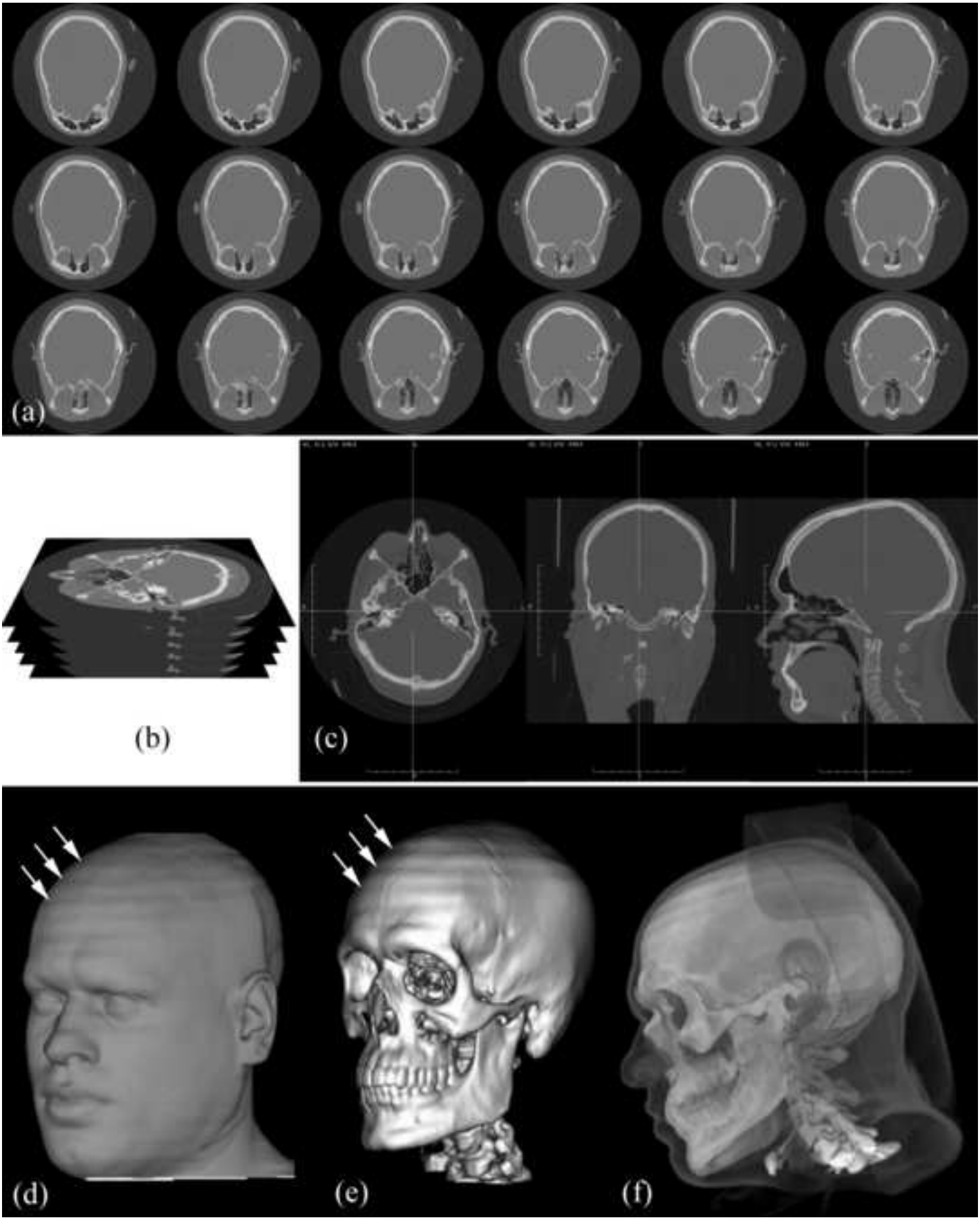

Figure 11

Click here to download high resolution image

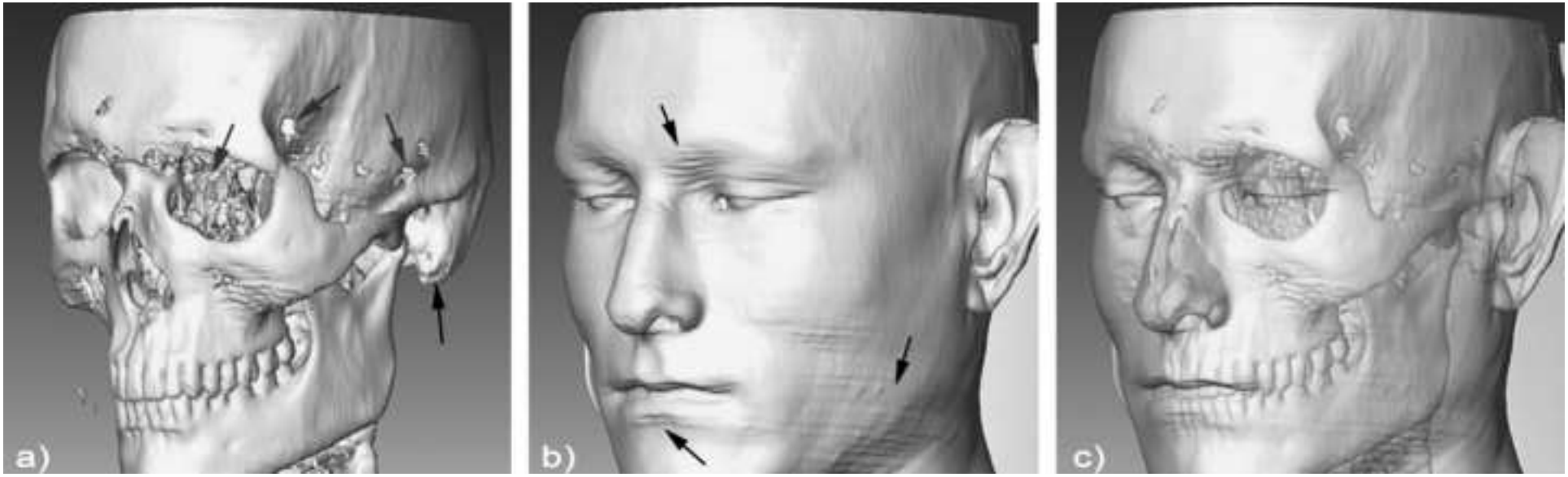

Figure 12

Click here to download high resolution image

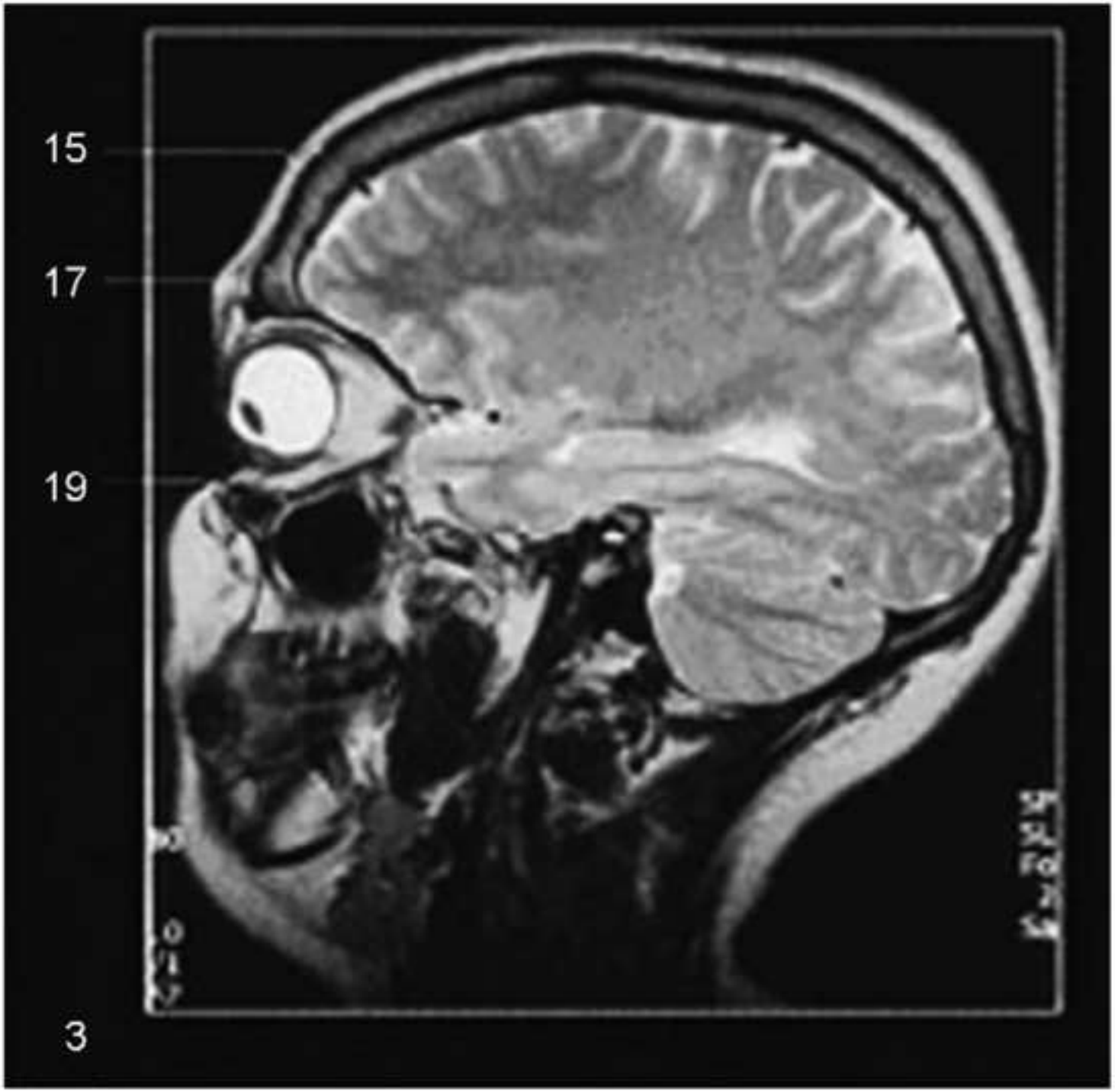

Table 1: Studies and sample sizes used to generate Figure 3.

		Needle Puncture	**Ultrasound**	**MRI**	**CT/CBCT**	**Radiographic Cephalogram**
Female FSTT	# of Studies	12	8	5	8	7
	Reference numbers	[57, 59, 62-66, 68, 80-83]	[16, 23, 70, 84-88]	[64, 89-92]	[22, 45, 48, 51, 52, 93-95]	[32, 34, 38, 60, 96-98]
	Pooled n	241	1564	288	552	709
Male FSTT	# of Studies	17	8	5	6	10
	Reference numbers	[57, 59, 62-66, 68, 80-83, 99-103]	[16, 23, 70, 84-88]	[64, 89-92]	[22, 46, 48, 51, 52, 94]	[19, 32, 34, 38, 60, 96, 97, 104-106]
	Pooled n	480	1280	324	413	822

Table 2: Error metrics reported for needle puncture studies using cadavers

Author	Date	Number of Remeasured Subjects	% of Sample Remeasured	Total Sample Size (N)	# Landmarks Used	Timeframe of Remeas-urement	Error Metric	Intra-observer Error	Inter-observer Error
Simpson & Henneberg [65]	2002	9	29	31	20	NR	TEM, rTEM and r	TEM (mm): mean = 1.4 , min = 0.2 , max = 3.0 rTEM (%): mean = 14.9 , min = 3.9 , max = 33.8 r: mean = 0.71, min = 0.17 , max = 0.96	TEM: mean = 1.6, min = 0.3 , max = 6.4 rTEM: mean = 12.8 , min = 5.2 , max = 33.8 r: no r provided
Domaracki & Stephan [67]	2006	10	30	33	10	NR	EM(CVE) and rTEM	TEM (mm): mean = 0.9, min = 0.4 , max =2.4 rTEM (%): mean = 9.7 , min = 3.4 , max = 24.1	NA
Suazo et al. [109]	2008	45	100	45	14	NR	k	k=0.93	k=0.93
Codinha [80]	2009	20	13	151	20	2hrs	MAE error and rTEM	MAE (mm): mean = 0.2, min = 0.1, max = 0.3 rTEM (%): mean = 1.6 , min = 1.0 , max = 2.5	NA

NR = not reported

Table 3: Error metrics reported for ultrasound studies using living participants

Author	Date	Number of Remeasured Subjects	% of Sample Remeasured	Total Sample Size (N)	# Landmarks Used	Timeframe of Remeas-urement (days)	Error Metric	Intra-observer Error	Inter-observer Error
Hodson et al. [122]	1985	5	10	50	20	intra-observer error: 17-35 Inter-observer error: <1	t-tests	No intra results reported	NS at 19 landmarks, one significant.

De Greef et al. [73]	2005	33	100	33	52	2-61	t-tests & Wilcoxon signed rank tests	Signed raw error = 0.09mm, sd = 2.17 mm; NS at 49 of 52 landmarks		NA
Ballie et al. [123]	2015	5	17	30	34	42	ICC using 2-way mixed model, absolute agreement and CI=95%	27 landmarks = 0.61-1.00 and 5 landmarks (0.41-0.60) (erect infraM2tooth R, Supine infraM2ridge R, Supine InfraM2ridge R, Erect InfraM2ridge L and Supine gonion L) dropped two landmarks due to negative		NA
Jia et al. [86]	2016	5	4	135	19	NR	NR	error not exceeding 0.5 mm		NA
Stephan & Preisler [87]	2018	8	13	63	6	Intra-observer error: >=2 Inter-observer error: <1	TEM, rTEM	TEM (mm): mean = 0.8 , min = 0.5 , max = 1.1 rTEM (%): mean = 10 , min = 9 , max = 12		TEM (mm): mean = 0.9 , min = 0.5 , max = 1.1 rTEM (%): mean = 11, min = 7, max = 19
Stephan & Sievwright [88]	2018	6	8	71	20	Intra-observer error: 13-126 Inter-observer error: 0-71	TEM, rTEM, r	TEM(mm): mean = 0.8 , min = 0.1 , max = 1.7 rTEM (%): mean = 8, min = 3, max = 15 r: mean = 0.73; min = -0.23, max = 0.97		TEM (mm): mean = 2.1, min = 0.4, max = 5.6 rTEM (%): mean = 21, min = 6, max = 45 r: mean = 0.29; min = -0.14, max = 0.78

NR = not reported

Table 4: Error metrics reported for CBCT & CT studies using living participants

Author	Date	Number of Remeasured Subjects	% of Sample Remeasured	Total Sample Size (N)	# Landmarks Used	Timeframe of Remeas-urement (days)	Error Metric	Intra-observer Error	Inter-observer Error
Kim et al. [44]	2005	1	100	1	6	28	ICC*	ICC = 0.996	ICC = 0.9960
Cavanagh & Steyn [45]	2011	22	14	154	28	NR	ICC*	min = 0.425, max = 0.919	min = 0.545, max = 0.946
Hwang et al. [46]	2012	20	20	100	31	28	TEM*	mean = 0.31, min = 0.07, max = 1.1	NA
Paneková et al. [47]	2012	20	13	160	14	14	absolute diff. and Mann-Whitney U Tests*	No significance, Min diff = 0, max diff = 5.3 mm, no means given	NA
Dong et al. [48]	2012	20	10	200	20	NR	r and TEM*	NA	r: mean = 0.81 , min = 0.60 , max = 0.92 TEM: mean = 0.93, min = 0.24 , max = 2.11
Perlaza Ruiz & Alonso [49]	2013	5	17	30	17	NR	t-test and pearson r*	r = 1.00, statistical significance unreported.	r = 0.99, statistical significance unreported.
Guyomarc'h et al. [50]	2013	20	5	500	37	28	CVE*	CVE less than 5%	CVE less than 5%
Bulut et al. [51]	2014	30	9	320	31	21	TEM, Mann-Whitney U test*	Mann-Whitney U test = no significant differences TEM: Less than 0.5mm for all landmarks (min 0.07 to max 0.40), mean = 0.20 mm	NA
Drgáčová et al. [52]	2016	5	5	102	40	NR	mean error*	mean error = 0.78 mm (not specified if intra- or inter-observer error)	
Lodha et al. [53]	2016	489	100	489	25	NR	one-way ANOVA*	no significance detected	NA
Thiemann et al. [54]	2017	24	8	320	38	28	r, TEM, rTEM*	r range = 0.94-0.97 TEM (mm) range = 0.03-0.23 rTEM (%) range = 0.15-1.92	r range = 0.96-1.00 TEM (mm) range = 0.03-0.49 rTEM (%) range = 0.17-3.91
Toneva et al. [55]	2018	15	20	75	16	7	ICC, TEM*	ICC > 0.98 TEM (mm) < max. 0.41	NA

* only single images were remeasured. Repeat images were not taken.

NR = not reported

Table 5: Error metrics reported for lateral radiograph studies using living participants

Author	Date	Number of Remeasured Subjects	% of Sample Remeasured	Total Sample Size (N)	# Landmarks Used	Timeframe of Remeas-urement (days)	Error Metric	Intra-observer Error	Inter-observer Error
Dumont [29]	1986	18	19	94	9	1	t-test*	no significance detected	NR
Genecov et al. [30]	1990	10	16	64	NR	NR	t-tests*	no significance detected	NR
Kasai [31]	1998	20	7	297	4	28	mean error*	mean error <1.0mm	NR
Garlie & Saunders [32]	1999	20	22	90	14	28	t-test*	NR	NR
Williamson et al. [33]	2002	20	9	224	13	1	t-tests*	no significance detected	no significance detected
Kurkcuoglu et al. [34]	2011	200	100	200	10	NR	rTEM*	min = 1.1 %, max = 8.6 %	NA
Fernandes et al. [35]	2013	20	19	105	10	30	TEM, t-test (systematic error)	2 out of 10 measurements significant TEM: mean = 0.62 , min = 0.31, max = 0.97	NR
Briers et al. [36]	2015	27	7	388	10	NR	ICC*	min = 0.99, max = 1.00	
Jeelani et al. [37]	2015	30	2	1357	11	21	ICC*	mean = 0.95, min = 0.83, max = 0.95	NR**
Wang et al. [38]	2016	50	20	256	10	7	r*	min 0.96 to max 1.00	min 0.92 to max 1.00
Kotrashetti & Mallapur [39]	2016	20	6	308	23	NR	paired t-tests*	only one test significant (gonion)	NR
Gibelli et al. [40]	2016	222	100	222	14	NR	t-tests*	no significance detected	no significance detected
Jeelani et al. [41]	2017	30	13	231	11	21	ICC*	mean = 0.96 , min = 0.87 , max = 0.99	NR**
Buyuk et al. [42]	2018	16	20	80	10	28	ICC*	>0.98	NR
Ayoub et al. [43]	2018	30	14	222	10	14	TEM*	min = 0.12, max = 0.76	min = 0.24, max = 0.86

* only single images were remeasured. Repeat images were not taken.

** study used multiple investigators but did not report inter-observer errors.

NR = not reported