ARTICLE INFO

Keywords:
Bone strength
Finite element analysis
Phantomless calibration
Internal tissue calibration
Bone mineral density
Hip fractures
Osteoporosis

ABSTRACT

Osteoporosis affects an increasing number of people every year and patient specific finite element analysis of the femur has been proposed to identify patients that could benefit from preventative treatment. The aim of this study was to demonstrate, verify, and validate an objective process for selecting tissues for use as the basis of phantomless calibration to enable patient specific finite element analysis derived hip fracture risk prediction. Retrospective reanalysis of patient computed tomography (CT) scans has the potential to yield insights into more accurate prediction of osteoporotic fracture. Bone mineral density (BMD) specific calibration scans are not typically captured during routine clinical practice. Tissue-based BMD calibration can therefore empower the retrospective study of patient CT scans captured during routine clinical practice. Together the method for selecting tissues as the basis for phantomless calibration coupled with the post-processing steps for deriving a calibration equation using the selected tissues provide an estimation of quantitative equivalent density results derived using calibration phantoms. Patient tissues from a retrospective cohort of 211 patients were evaluated. The best phantomless calibration resulted in a femoral strength (FS) [N] bias of 0.069 ± 0.007% over FS derived from inline calibration and a BMD [kg/cm²] bias of 0.0038 ± 0.0037% over BMD derived from inline calibration. The phantomless calibration slope for the best method presented was within the range of patient specific calibration curves available for comparison and demonstrated a small bias of 0.038 ± 0.037 HU/(mg/cm³), assuming the Mindways Model 3 BMD inline calibration phantom as the gold standard. The presented method of estimating a calibration equation from tissues showed promise for CT-based femoral fracture analyses of retrospective cohorts without readily available calibration data.

1. Introduction

Over 300,000 people experience an osteoporotic femoral fracture in the U.S. every year [1]. Despite available treatments, osteoporosis remains underdiagnosed [2], inspiring research towards a better understanding of osteoporotic fracture. In addition, the stratification accuracy of the prognostic standard of care (bone densitometry) is too low to reliably diagnose osteopenic patients, and to decide when to adopt second-line treatments such as Denosumab or Teriparatide [2,3]. This calls for more accurate prognostic methodologies. Various groups proposed quantitative computed tomography (QCT) based patient specific finite element analyses (FEAs) for improved osteoporotic hip fracture risk assessment [4–6]. These FEAs have been shown to predict risk of hip fracture more accurately than areal bone mineral density (BMD) [4]. Retrospective reanalysis of patient computed tomography (CT) scans will further assist in the development of techniques to predict risk of osteoporotic fracture, potentially leading to improved prognostic accuracy. However, these models depend on the estimation of bone material properties, derived from CT X-ray attenuation. In phantom-based calibration this is achieved by placing an inline calibration phantom under the patient or by scanning offline a calibration phantom immediately after the patient, using the same CT scan settings. Phantom-based calibration is the gold standard in the development of patient specific FEAs.
However, scanning the patient with an inline phantom is not a standard clinical practice, and delayed offline retrospective calibration is not always possible due to clinics regularly purchasing new CT scanners. Phantomless CT scan calibration can be derived from patient tissues and could therefore be a feasible alternative.

Before considering literature on existing phantomless methods, several variables should be identified and defined. There are several points in the process of capturing a CT scan that affect density assessment including: underlying theory and definitions, the chemical composition of the object being scanned, the acquisition settings and the reconstruction algorithms. Considering underlying theory, clinical CT images describe materials’ X-ray attenuation in greyscale in terms of the Hounsfield Scale (in units HU),

\[
CT \text{ Number} = \left( \frac{\mu (E) - \mu_{\text{water}}}{\mu_{\text{water}} - \mu_{\text{air}}} \right) \times 1000 \ [\text{HU}]
\] (1)

where \(\mu\) is X-ray attenuation from the object, represents \(\mu(E) = a_1 \text{PE}(E) + a_2 \text{CS}(E) = m_1 \mu_1(E) + m_2 \mu_2(E)\) (2)

For CT Numbers to be used in clinical research, the calibration slope than HA at equivalent densities \([26,27]\), underestimating cortical bone density. They specifically pointed out that this difference is more pronounced at higher densities, visually demonstrated by Knowles et al. \([10]\). Phantomless calibration, by definition, removes the variations created by scanning a phantom, retains the potential to create a scan specific calibration equation and increases accuracy over an inline phantom by moving the density reference closer to the bone.

To enable density assessment of patient scans where phantom-based calibration data were not captured, three approaches to phantomless calibration have been used in clinical research \([28]\): (1) using CT Numbers \([\text{HU}]\) directly \([29–32]\); (2) using a calibration factor \([21,23,33,34]\); and, (3) substituting tissues as a calibration reference \([15,16,18–23,26,31–38]\). The first approach, using CT Numbers \([\text{HU}]\) directly, is most accessible within current clinical practice limitations. Unfortunately, in order to be considered quantitative, the relevant BMD thresholds would have to be CT scanner and protocol specific. Trying to derive relevant FEAs based thresholds in terms of CT Numbers would pose challenges, such as requiring significant amounts of patient case studies. In the second approach, using a calibration factor, a general calibration factor \((GCF)\) is calculated as the ratio of \(QCT\) derived BMD divided by CT Numbers \([\text{HU}]\) and then rearranged to extrapolate phantomless BMD through multiplying CT Number \([\text{HU}]\) by GCF \([21]\). While this approach is CT scanner and protocol specific, it is neither scan specific nor precise enough for FEAs. The third approach, substituting tissues as calibration references is scan specific, and has been applied in

\[
\rho_{\text{HA}} = \left( \frac{\text{CT Number} - b}{m} \right)
\]

(3)

where \(m\) \([\text{HU}/(\text{mg/cm}^3)]\) and \(b\) \([\text{HU}]\) are the slope and intercept, respectively, from the linear regression. When density reference phantoms are used, the derivation of the calibration equation naturally characterizes and accounts for CT number variations due to factors including manufacturer, model and protocol \([24]\). The use of standardized and stable references in modern density phantoms can provide a comparison for analyses across clinics. However, in the case of an inline phantom that is externally located under the patient, the phantom will be subjected to patient-modified spectra variable with patient composition, size, and geometric position \([25]\). While scanning an offline phantom removes this variation, this calibration method does not capture differences such as those created by dosage-reducing variable current algorithms. Initially intended to create a standardized reference to characterize variations in CT number, differences in phantoms now introduce additional variances and limitations into the comparison of clinical assessment techniques. For example, \(K_2\text{HPO}_4\) in place of HA was discussed by Cann et al. who argued \(K_2\text{HPO}_4\) results in a slightly lower calibration slope than HA at equivalent densities \([26,27]\), underestimating cortical bone density. They specifically pointed out that this difference is more pronounced at higher densities, visually demonstrated by Knowles et al. \([10]\).
FEAs of the femur [17,22,36,37,42]. This method is limited by the assumption that internal patient tissues have specific densities [42]. Previously, a variety of tissues served as the basis for deriving phantomless calibration: fat and muscle [19,20,25,35,39–41]; air and blood [17,36,37]; air and fat [17,36,37]; air, fat, and muscle [22]; and air, fat, blood, muscle, and cortical bone [22,42]. Many things are known to influence the ability of CT Numbers (HU) to measure tissues: hydration levels [20], patient pathologies [43], heterogeneous distributions of muscle and fat [20], and i.v. contrast [19,44]. Further, CT is unable to assess some pathologies known to affect CT Number, such as fatty atrophy of muscle [39].

While there is no standard method for determining which tissues to use as the basis for phantomless calibration, the literature provides some rationale for choosing specific tissues. Boden et al. showed fat and muscle offer reliable internal reference standards for measuring vertebral bone density with QCT using tabulated reference densities from White [25,45]. More recently, Michalski et al. used tabulated and standardized mass attenuation coefficients from the National Institute of Standards and Technology (NIST) [42,46]. Some researchers have attempted to determine their own ground truth values using a system of equations approach, finding: –69 mg/cm$^3$ for fat and 77 mg/cm$^3$ for muscle [40]; or –840 mg/cm$^3$ for air, –80 mg/cm$^3$ for fat, and 30 mg/cm$^3$ for muscle [22]. The limitation to deriving ground truth values, in lieu of using the standardized tables, is the unknown amount of pathological variation in the base cohort.

In the absence of phantom-based calibration data, computational researchers commonly estimate a linear relationship between a specific density and CT Number based on available literature. Two such densities include ash density, ash mass divided by bulk sample volume, and apparent density, wet mass without marrow divided by bulk sample volume [10]. Several studies are available in literature where bench researchers empirically derived linear relationships between either ash density or apparent density measurements of bone and CT number [47–54]. Ford et al. demonstrated a method for estimating a linear relationship between apparent density and CT Number for trabecular bone and cortical bone in mg/cm$^3$ [40], using Eq. (6), and finally Young’s modulus using Eq. (7) [53].

\[ \rho_{\text{app}} = 1.06 \times \text{CT Number} + 68.4 \quad (4) \]

before using the relationship in a computational study [55]. Though not demonstrated in literature, another approach would be to estimate soft tissue density by estimating a theoretical calibration slope (CT\text{theoretical} = 1.025 \text{HU/(mg/cm}^3\text{)}) derived from theoretical air (1.205 mg/cm$^3$, –1024 HU) and theoretical water (1000 mg/cm$^3$, 0 HU). Both of these density estimation methods do not take into consideration CT scanner performance parameters or the anatomical area as phantom-based or tissue-based phantomless calibration estimates do.

The method for deriving Young’s modulus, a measurement of material stiffness, from CT data for use in patient specific FEAs is sensitive to the relationship between a specific density and CT Number due to a power-law relationship [10]. CT data are converted to \( \rho_{\text{aph}} \) using Eq. (5) depending on the equivalent density, then \( \rho_{\text{app}} \) using Eq. (6), and finally Young’s modulus using Eq. (7) [53].

\[ \rho_{\text{aph}} = 0.8772 \times \rho_{\text{CT}} + 0.07895 \quad (5) \]

\[ \rho_{\text{app}} = 0.598 \times \rho_{\text{aph}} \quad (6) \]

\[ E = 6850 \times \rho_{\text{app}}^{0.49} \quad (7) \]

In addition to being specific to the phantom’s reference material, these relationships are also specific to anatomic site, in this case the femur [56]. This suggests a need for a method flexible enough to consider anatomic site when selecting reference tissues for phantomless calibration.

The aim of this retrospective study was to demonstrate, verify and validate a method for selecting patient tissues from which to derive density for use in femur strength prediction. Using the selected tissue combinations, we present a method for using phantomless calibration to estimate bone material properties for later use in femoral fracture risk prediction. Using a $Z^2$-factorial design, we tested repeatability with and without theoretical data points, and with and without including multiple scans for each patient. For verification, we compared patient specific results against an offline custom CIRS BMD phantom, and an inline Mindways Model 3 BMD calibration phantom. For validation, we compared patient specific results against the inline Mindways Model 3 BMD phantom for the patients whose scans included the phantom.

2. Materials and methods

Patient scans were selected for a density-related sensitivity analysis from data gathered previously related to a cohort of 408 patients gathered at the University of Wisconsin-Madison hospital. Scans from this cohort were previously identified to examine femoral fracture in an age matched, case-control study. Full details of the previous study are available in Lee et al. from 2017 [57]. Retrospective CT scan analysis was Health Insurance Portability and Accountability Act compliant and approved by the UW-Madison Institutional Review Board (protocol number 2016-0168).

The pre-fracture cases analysed consisted of 43 patients, with 26 female patients (ages 50–93 years) and 16 male patients (ages 56–95 years). The average time to fracture after CT scan was 1 year, with the minimum occurring the same year and the maximum occurring within 4 years. The control cases analysed consisted of 168 patients, with 108 female patients (ages 50–90 years) and 60 male patients (ages 50–91 years).

2.1. Method of selecting patient scans for analysis

Scans analysed were limited to scans captured on a GE Lightspeed family CT scanner including: LightSpeed 16, LightSpeed Pro 16, Light-Speed Pro 32, LightSpeed Ultra, LightSpeed VCT, Discovery CT750HD, Optima 580, Optima 660 HD, and Revolution GSI. All scans analysed were captured at 120 kVp, and 1.25 mm slice thickness. The 258 scans analysed (Table 1) included images of 211 individual patients, both male and female (aged 50 to 95 years). Patients with surgical hardware were excluded from the study. Our goal in this selection was to cover a broad range of data for the phantomless calibration to be broadly applicable so we processed all data that met our inclusion criteria.

2.2. CT scanning protocol

Images were collected during routine abdominopelvic CT scans performed using 16- to 64-Multi-Detector CT scanners (LightSpeed Series, GE Healthcare). Hospital routine includes daily calibration scans on each machine to ensure the accuracy of the CT attenuation values. Standard scanning parameters for routine abdominopelvic CT scans are 120 kVp tube voltage, 1.25 mm slice thickness, 0.625 mm slice spacing, a medium or body type filter, a standard convolution kernel, and low doses of current, either static (50–100 mA) or modulated (noise index, 50; range 30–300 mA).

2.3. Inline quantitative equivalent density calibration using the Mindways model 3 BMD calibration phantom

Eight out of the 408 patient scans included an inline effective K2HPO4 density calibration phantom (Model 3 Phantom, Mindways Software, Inc., Austin, TX). Of those eight, three patients had existing surgical hardware and could not be analysed. Therefore, the analyses in this paper were limited to five patients. The calibration process for this phantom is described in detail by Mindways [58]. Manual calculation of the calibration slopes for the five patients scanned with the inline calibration phantom was conducted (Table 2). A power analysis for a two-
sample pooled t-test was conducted in MATLAB and the necessary sample size to meet 99% power ranged between 2 and 5 for the majority of the 40 phantomless slope combinations considered, with 3 outliers requiring a sample size of 8.

2.4. Offline equivalent density calibration using a custom BMD phantom

Retrospectively, we scanned offline a custom phantom with four HA density plugs at 100, 400, 1000 (part: 06217), and 1750 (part: 06221) mg/cm\(^3\) (CBRS Inc., Norfolk, VA) submerged in water. Scan settings were 120 kVp, 1.25 mm slice thickness, 0.625 mm slice spacing, 100 mA, and a standard reconstruction kernel on the Discovery 750HD. HA plug densities were selected to be representative of human femoral bone [11]. Plugs were segmented by creating a virtual cylinder with a 10-

\[
\rho = 0.9091 \times CT Number - 23.90
\]

2.5. Identify most consistent reference densities across patients

We analysed phantomless calibration on 258 scans and considered five patients’ nominal density references, including adipose tissue, aortic blood, skeletal muscle, urine, and air. Tissue segmentations were captured as virtual cylinders, with a diameter of 10 pixels and a depth of 10 slices, using Mimes v21.0 (Materialise, Leuven, Belgium). Due to the small size of the femoral artery, the virtual cylinder captured was reduced to a diameter of 8 pixels. For consistency, all virtual cylinders were created such that the centre of the virtual cylinder was around the same axial slices as the centre of the femoral head. An example of the virtual cylinder placement is shown in Fig. 1. Quality checks were conducted to ensure each virtual cylinder contained a volume of at least 100 voxels (ASTM E1935 2019). We were unable to segment urine in the patient’s bladder for 167 out of the 258 scans due to empty bladders. Aortic blood was also difficult to segment due to their small sizes, resulting in measured values outside of 40 ± 20 HU for 46/258 left patient arteries and 47/258 right patient arteries. Table 3 shows the nominal density values assumed for the linear regression of HU and tissue density [46]. The 258 patients included in this study were segmented by a single operator. To assess the precision of results at the segmentation, BMD and FEAs levels, the five patients with inline phantoms were also segmented by three different operators. Inter- and intra-operator repeatability was calculated [59].

Each patient had up to nine potential data points that could be used for line fitting: theoretical air, segmented air, adipose tissue (right and left), aortic blood (right and left), skeletal muscle (right and left), and theoretical water. Any combination of at least two and up to nine data points could be used to derive a linear regression for the HU versus nominal density relationship, 502 possible combinations for each of the 258 scans. A custom MATLAB (v.2018b, The MathWorks, Inc., Natick, MA, US) script was developed to: (1) calculate all possible linear regressions, (2) discard all ill-conditioned calibration slope results, and (3) conduct a numerical analysis to sort density combination calibration slope results across patients. Ill-conditioned calibration slopes occurred when the algorithm fit a line with two values for the same tissue (i.e. right and left adipose). Sorting was accomplished by minimizing the sum of the squared error between the density calibration slope and a theoretical calibration slope,

\[
\sum (m - 1.025)^2
\]

Recall from the introduction that the theoretical calibration slope (1.025 HU/(mg/cm\(^3\))) is derived from theoretical air (1.205 mg/cm\(^3\), -1024 HU) and theoretical water (1000 mg/cm\(^3\), 0 HU). After discarding over-constrained combinations, the best 10 combinations and the worst combination were identified for further analysis.

2.6. Experimental design to test repeatability of tissue identification

Patient tissue segmentations were organized to form two groups: “Scans” included all scans eligible for processing for all patients, and “Patients” included only one scan for each patient. To form the Patients group, results from duplicate scans for patients were removed such that the results for CT scanners with fewer patient scans were kept except in the case of the Optima 580 and Revolution GSI, each of which only had

<table>
<thead>
<tr>
<th>Patient</th>
<th>CT scanner</th>
<th>Calibration slope [HU/(mg/cm(^3))]</th>
<th>Y-intercept [HU]</th>
<th>Conversion Slope [mg/cm(^3)]/HU</th>
<th>Y-intercept [mg/cm(^3)]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 LightSpeed VCT</td>
<td>1.03</td>
<td>5.59</td>
<td>0.97</td>
<td>-5.43</td>
<td></td>
</tr>
<tr>
<td>2 LightSpeed Pro 16</td>
<td>1.06</td>
<td>3.32</td>
<td>0.94</td>
<td>-3.12</td>
<td></td>
</tr>
<tr>
<td>3 LightSpeed</td>
<td>1.05</td>
<td>13.54</td>
<td>0.95</td>
<td>-12.88</td>
<td></td>
</tr>
<tr>
<td>4 LightSpeed Pro 16</td>
<td>0.99</td>
<td>5.87</td>
<td>1.01</td>
<td>-5.93</td>
<td></td>
</tr>
<tr>
<td>5 Discovery CT750HD</td>
<td>1.06</td>
<td>-8.5</td>
<td>0.94</td>
<td>8.00</td>
<td></td>
</tr>
</tbody>
</table>

Average values 1.04 | 3.96 | 0.96 | -3.87 |
one patient scan. A $2^2$ factorial designed experiment was conducted by running the MATLAB script used to identify the most consistent reference densities across patient populations two levels for each group including and excluding values for theoretical air and water in the combinatorial analysis.

2.7. Finite element model BMD and femur strength

Five finite element models were developed for each patient to investigate the impact of different calibration equations on BMD and femoral strength (FS) calculations model I: patient specific inline $K_2HPO_4$ calibration; model II: the average of the patient specific inline $K_2HPO_4$ calibrations; model III: the offline HA calibration; model IV: phantomless calibration derived from air, aortic blood, and skeletal muscle (AABSM); and, model V: phantomless calibration derived from air and adipose (AA). One femur was segmented for each patient: four were segmented in Mimics v19.0 or 21 (Materialise, Leuven, Belgium) and one was segmented in ITK-Snap (ITK-Snap 3.6.0, University of Pennsylvania). Each geometry was discretized into ten-node tetrahedral elements using ICEM CFD (ICEM CFD 16.2, Ansys Inc., PA, USA) with a maximum edge length of 3 mm based on a previous mesh convergence study [60]. Note that each patient had the same mesh for all models.

Elastic moduli were mapped onto the meshed bone using the equations described in the introduction and Bonemat (V3.2, Istituto Ortopedico Rizzoli, Bologna, Italy). BMD was calculated for each model as the summation across groups of the density in each material group multiplied by the number of elements with that material group. Femur strength was calculated using a sideways fall loading scenario with fixed boundary constraints at the estimated knee centre and a simulated planar bearing at the lateral coordinate on the trochanter [4, 61]. A concentrated point load, 1000 N, was applied to the centre of the femoral head in thirty-three different force directions from $-30^\circ$ to $30^\circ$ (posteriorly to anteriorly directed) in the transverse plane and $0^\circ$ to $30^\circ$ (x-axis to medially directed) in the frontal plane [61]. FEAs strain results were post-analysed using a maximum principal strain failure criterion with limiting values at 0.73% for tensile and 1.04% for compressive strains as previously defined by Bayraktar et al. [62]. FS was defined as the minimum force (N) at failure across all 33 side-fall loading conditions. All FEAs were conducted in ANSYS 16.2 (Ansys Inc., PA, USA).

2.8. Statistical analysis

The mean and standard deviation were calculated for patient tissue segmentation measurement results in HU for both the “Scans” and “Patients” groups. Once patient specific density calibration slopes were calculated, statistical measurements were mean, standard deviation, and 95% confidence interval. Bland-Altman analyses were conducted for the five patients with inline Mindways Model 3 BMD phantoms included in

<table>
<thead>
<tr>
<th>Reference material</th>
<th>Nominal density [mg/cm$^3$]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Theoretical air (not segmented)</td>
<td>1.205</td>
</tr>
<tr>
<td>Adipose tissue</td>
<td>950</td>
</tr>
<tr>
<td>Aortic blood</td>
<td>1060</td>
</tr>
<tr>
<td>Theoretical water (not segmented)</td>
<td>1000</td>
</tr>
<tr>
<td>Skeletal muscle</td>
<td>1050</td>
</tr>
</tbody>
</table>

Table 3 Nominal density values from NIST [46].

Fig. 1. Representative axial slice of a CT scan of an 85-year-old patient at the proximal femur level. The patient is lying supine. A Mindways Model 3 BMD Calibration Phantom is visible underneath the patient.
their scans. The hypotheses that no statistically significant difference exists between calibration methods were tested using a Students’ t-test ($\alpha = 0.01$) for the calibration slopes, the calibration intercepts, BMD, and FS. Normality was tested using the Shapiro-Wilk test in IBM SPSS Statistics for Windows, version 26 (IBM Corp., Armonk, N.Y., USA); however, all other statistics were calculated in MATLAB 2018b.

3. Results

Phantomless calibration was valid when compared against inline phantom calibration for FS, BMD, calibration equation (Figs. 2, 3, 4). The algorithm produced calibration equation results consistent with those from inline phantom calibration (Fig. 5). Intra- and inter-operator repeatability found the method highly repeatable for FS, BMD, and calibration equations (Table 4). Of the tissues segmented, adipose was the most repeatable and the bladder was the least repeatable (Table 4).

The AABSM combination produced the best slope result for 3 of the 4 categories in the 2$^\text{nd}$ factorial designed experiment. The 4th category, excluding multiple scans per patient and theoretical air and water, found the AA combination produced the best slope. The first category, including theoretical air and theoretical water for all scans ($n = 258$), found AABSM scan specific slope values (HU/(mg/cm$^2$)) of mean ± std. dev (lower–upper) = 1.021 ± 0.006 (1.008–1.034) and found measured air and theoretical water produced the worst combination with slope values of 1.379 ± 6.185 (–10.99–13.75). The second category, including theoretical air and theoretical water for 1 scan per patient ($n = 211$), found AABSM scan specific slope values of mean ± std. dev (lower–upper) = 1.021 ± 0.006 (1.009–1.034) for the best combination and found measured air and theoretical water produced the worst combination with slope values of 1.468 ± 6.856 (–12.24–15.18). The third category, excluding theoretical air and theoretical water for all scans ($n = 258$), found AABSM scan specific slope values of mean ± std. dev (lower–upper) = 1.017 ± 0.010 (0.998–1.037) for the best combination and found aortic blood and skeletal muscle produced the worst result with values of 0.893 ± 2.151 (–3.458–5.195). The final category, excluding theoretical air and water for 1 scan per patient ($n = 211$), found AA scan specific slope values of mean ± std. dev (lower–upper) = 0.975 ± 0.010 (0.956–0.994) and found aortic blood and skeletal muscle produced the worst result with values of 0.839 ± 2.149 (–3.458–5.137).

For FS results, the AABSM calibration resulted in a 6.9% bias over scan specific inline calibration, a 7.3% bias over averaged inline calibration, and a 22% bias over offline calibration; and the AA calibration resulted in a 9.9% bias over scan specific inline calibration, a 10% bias over averaged inline calibration, and a 25% bias over offline calibration (Fig. 2). For BMD results, the AABSM calibration resulted in a 3.9% bias over scan specific inline calibration, a 3.7% bias over averaged inline calibration, and a 17% bias over offline calibration; and the AA calibration resulted in a 6.1% bias over scan specific inline calibration, a 6.0% bias over averaged inline calibration, and a 19% bias over offline calibration (Fig. 3). When considering the calibration slopes directly, the AABSM and AA combinations resulted in biases of 2.6% and 6.3% over scan specific inline calibration, respectively (Fig. 4). For the calibration

<table>
<thead>
<tr>
<th>FS Validation</th>
<th>Air, Artery, Skeletal Muscle</th>
<th>Air, Adipose</th>
</tr>
</thead>
<tbody>
<tr>
<td>K$_2$HPO$_4$ equivalent BMD</td>
<td><img src="image1" alt="Graph" /></td>
<td><img src="image2" alt="Graph" /></td>
</tr>
<tr>
<td>Phantomless - QCT/QCT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average [N]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>K$_2$HPO$_4$ equivalent BMD</td>
<td><img src="image3" alt="Graph" /></td>
<td><img src="image4" alt="Graph" /></td>
</tr>
<tr>
<td>Phantomless - QCT/QCT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average [N]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HA equivalent BMD</td>
<td><img src="image5" alt="Graph" /></td>
<td><img src="image6" alt="Graph" /></td>
</tr>
<tr>
<td>Phantomless - CT/CT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average [N]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fig. 2. FS results derived using phantomless calibration displayed the least bias when compared against results derived using the average of the patient and scan-specific K$_2$HPO$_4$ calibration as shown by Bland-Altman analyses. Overall results using phantomless calibration were more consistent with results from the K$_2$HPO$_4$ phantom than the HA phantom. The blue lines are the means and the red lines are the 95% confidence interval. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)
When comparing scan specific results for all 211 patient scans against results derived using the average of the patient and scan specific \( K_2 HPO_4 \) calibration as shown by Bland-Altman analyses. Overall results using phantomless calibration were more consistent with results from the \( K_2 HPO_4 \) phantom than the HA phantom. The blue lines are the means and the red lines are the 95% confidence interval. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Differences in FS between calibration methods were only statistically significant for AABSM versus the average of the inline calibrations and the offline phantom (\( p < 0.001, 0.001 \)). For the intercepts, differences were not statistically significant between either phantomless calibration combinations (AABSM and AA) versus the average of the inline calibrations (\( p < 0.001, 0.001 \)) and the offline phantom (\( p < 0.001, 0.001 \)). For the intercepts, differences were not statistically significant between either phantomless calibration combination (AABSM and AA) and the inline phantom (\( p = 0.04, 0.17 \)). Conversely, differences were statistically significant between both phantomless calibration combinations (AABSM and AA) versus the average of the inline calibrations (\( p < 0.001, 0.001 \)) and the offline phantom (\( p < 0.001, 0.001 \)). For the intercepts, differences were not statistically significant between either phantomless calibration combination (AABSM and AA) and the inline phantom (\( p = 0.08, 0.26 \)). Continuing with the trend, differences were statistically significant between both phantomless calibration combinations (AABSM and AA) versus the average of the inline calibrations (\( p < 0.001, 0.001 \)) and the offline phantom (\( p < 0.001, 0.001 \)).

Both average intra-operator and inter-operator repeatability were better for AABSM than for AA when analysing FS, BMD, or calibration equation (Table 4). Segmentation CT Number [HU] results found similar means and standard deviations for tissues compared between the “all scans” and “one scan per patient” categories, respectively: adipose 99.12 ± 9.44 and 98.98 ± 9.62; aortic blood 52.42 ± 17.28 and 51.83 ± 17.41; and muscle 43.58 ± 13.42 and 44.01 ± 14.05.

4. Discussion

The main aim of this study was to demonstrate, verify and validate a method for selecting basis patient tissues for deriving an equivalent density equation in femoral bone QCT analyses. As an example, this method identified AABSM as the best combination of tissues for phantomless calibration. This method was shown to be valid for FS, BMD, and calibration equation results. The validity of phantomless calibration for FEAs of the femur is consistent with other studies [17,37,42]. To verify...
this method, results for 258 scans were shown to be within the range of those from the inline calibration of five scans. This method shows promise for use in the retrospective analysis of patient cohorts without available calibration information and can be applied opportunistically to any CT scan.

This study differs from previous studies in several ways including different CT scanners, CT scan protocols, tissues used as the basis for phantomless calibration, assumed tissue densities, methods of segmentation, and FEA pipelines. Focusing in on which tissues are used as the basis for phantomless calibration, this study’s selection of the AABSM combination of tissues is different from prior combinations in literature for FEAs of the femur, including: fat and muscle [39]; air and fat

Fig. 4. Phantomless calibration slopes derived from air, aortic blood and skeletal muscle segmentations displayed less bias than those derived from air and adipose when compared with patient and scan specific $K_2HPO_4$ calibration as shown by Bland-Altman analyses. While both sets of phantomless calibration intercepts displayed similar and large bias, all averages were within the performance expectations for a GE CT scanner (0 ± 7 HU) [24]. The blue lines are the mean and the red lines are the 95% confidence interval. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Fig. 5. These plots compare the ten best combinations of tissues in terms of calibration slopes and intercepts. Boxplots are overlaid on scatter plots of the patient specific calibration slopes and intercepts (purple points). For the slopes plots, the three blue lines include: the dashed lines for the minimum (0.99 HU/(mg/cm$^3$)) and maximum (1.06 HU/(mg/cm$^3$)) slopes across patients from the $K_2HPO_4$ calibration phantom, and the dash-dot line is for the calibration slope for the custom phantom scanned offline in water (1.10 HU/(mg/cm$^3$)). All slopes are in HU/(mg/cm$^3$). For the intercept plots, the three blue lines include: the dashed lines for the minimum (−0.0085 HU) and maximum (0.0135 HU) patient specific results for the $K_2HPO_4$ phantom, and the dash-dot line is calibration intercept for the custom phantom scanned in water (−0.0239 HU). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)
Table 4
Intra- and Inter-operator reanalysis precision error (root-mean-square) for FS, BMD, calibration equation, and tissue segmentations at the femur for n = 5. Coefficients of variation (CV_RMS, in %) and standard deviations (SD_RMS, in absolute units) are presented.

<table>
<thead>
<tr>
<th>Measurement</th>
<th>Intra-operator</th>
<th>Inter-operator</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CV_RMS (%)</td>
<td>SD_RMS</td>
</tr>
<tr>
<td>Femoral strength [N]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AABSM</td>
<td>0.28</td>
<td>9.67</td>
</tr>
<tr>
<td>AA</td>
<td>1.27</td>
<td>63.85</td>
</tr>
<tr>
<td>Inline</td>
<td>0.37</td>
<td>11.68</td>
</tr>
<tr>
<td>Bone mineral density [kg/cm³]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AABSM</td>
<td>0.22</td>
<td>0.13</td>
</tr>
<tr>
<td>AA</td>
<td>0.65</td>
<td>0.48</td>
</tr>
<tr>
<td>Inline</td>
<td>0.57</td>
<td>0.57</td>
</tr>
<tr>
<td>Slope [HU/(mg/cm³)]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AABSM</td>
<td>0.22</td>
<td>0.00</td>
</tr>
<tr>
<td>AA</td>
<td>0.41</td>
<td>0.01</td>
</tr>
<tr>
<td>Inline</td>
<td>0.60</td>
<td>0.01</td>
</tr>
<tr>
<td>Intercept [HU]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AABSM</td>
<td>2.70</td>
<td>0.08</td>
</tr>
<tr>
<td>AA</td>
<td>6.20</td>
<td>7.45</td>
</tr>
<tr>
<td>Inline</td>
<td>9.31</td>
<td>2.22</td>
</tr>
<tr>
<td>Segment [HU]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adipose</td>
<td>1.68</td>
<td>3.77</td>
</tr>
<tr>
<td>Air</td>
<td>0.35</td>
<td>7.44</td>
</tr>
<tr>
<td>Aortic blood</td>
<td>10.09</td>
<td>10.67</td>
</tr>
<tr>
<td>Skeletal muscle</td>
<td>10.25</td>
<td>7.19</td>
</tr>
<tr>
<td>Bladder</td>
<td>50.82</td>
<td>3.67</td>
</tr>
</tbody>
</table>

Despite the differences in FEA pipelines, the bias introduced by phantomless calibration is comparable across studies, with all other variables held constant within the respective studies. This study’s calculated FS mean absolute difference, 90 N (6.9%), was similar to recent studies such as Lee et al. 30 N (0.8%) [17], and Michalski et al. –40 N (17%) [39]. The calculated BMD biases 0.92 kg/cm³ (0.04%) were larger than a recent study on a more developed method presented by Lee et al., 2 mg/cm³ (0.9%) [17]. Note that differences observed in FS measurements were expected to be greater than differences observed in BMD measurements for two reasons. First, differences that appear small when examining preliminary results (i.e. segmentation, calibration equation, BMD) are amplified by the power-law component of the density-elastic modulus relationship (Eq. (7)) making FEAs sensitive to changes in the calibration equation. Second, the side-ways fall load case is more sensitive to changes in mechanical properties of materials due to the stress gradient from bending in the combined-loading. Both the results of this study and the results from literature show greater differences in FS biases than BMD biases. From a clinical perspective, this drives the need for a universally accessible objective method, such as that presented in this study, for identifying the best tissues for use as the basis for FEA, allowing them to make relevant observations (1) changing reconstruction kernel had no significant effect on phantom-based or air-fat-muscle calibration, and (2) changing reconstruction kernel resulted in significantly higher failure loads when using their non-patient specific calibration [22]. Michalski et al. observed that by using consistent imaging acquisition and a single imaging protocol there were fewer confounding variables when measuring methodological precision [42]. Beyond the limitation of only considering one clinical protocol, this study was also limited to pre-fracture cases that went on to experience femoral fragility fracture.

The current study’s segmentation method may be less repeatable than the segmentation methods presented in other studies. Where this study conducted manual segmentation using the mean CT Number [HU] over the digital volume, other studies used higher fidelity segmentation methods. Examples relevant to multiple studies include: Lee et al. who have automated their segmentation using gradient-profile algorithms independent of absolute attenuation [17,36,37], and the popular histogram and peak fitting approach [20,22,25,41,42]. Boden et al. designed the histogram and peak fitting approach specifically to overcome the challenge of reliably locating a conventional ROI to calculate the mean CT Number [HU] of the digital volume [25]. This implies that methods using this approach would naturally account for the heterogeneity included in patient tissues and improving the precision of phantomless calibration. The differences in segmentation methods are a major reason why this method was less repeatable than those presented previously in literature (Table 5). This comparison shows using a higher fidelity segmentation method may improve the repeatability of the current study’s phantomless calibration method.

This study showed phantomless calibration results were close to results derived from the Midways Model 3 BMD inline phantom which relies on K₂HPO₄ as a reference material. Further, the phantomless calibration derived results were not significantly different from the inline calibration derived results and were significantly different from both the averaged inline calibration and the offline calibration. Both the inline phantom, which ranges from –53.4 to 375.8 of equivalent K₂HPO₄, and this phantomless calibration technique, require extrapolation in order to define in vivo BMD [58]. The potential for extrapolation errors has been raised as a concern in several studies [20,22,41]. In their phantomless study, Lee et al. demonstrated their method to

Table 5
Comparison of precision errors between the current study and literature.

<table>
<thead>
<tr>
<th>Measurement</th>
<th>Intra-operator</th>
<th>Inter-operator</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CV_RMS (%)</td>
<td>SD_RMS</td>
</tr>
<tr>
<td>Femoral strength [N]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lee [17] [N]</td>
<td>0.4</td>
<td>20</td>
</tr>
<tr>
<td>Michalski [42] [N]</td>
<td>6.0</td>
<td>84</td>
</tr>
<tr>
<td>Current study [N]</td>
<td>0.28</td>
<td>9.67</td>
</tr>
<tr>
<td>Bone mineral density [mg/cm³]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lee [17] [mg/cm³]</td>
<td>&lt;0.3</td>
<td>1</td>
</tr>
<tr>
<td>Michalski [42] [mg/cm³]</td>
<td>4.3</td>
<td>12</td>
</tr>
<tr>
<td>Current study [kg/cm³]</td>
<td>0.22</td>
<td>0.13</td>
</tr>
</tbody>
</table>
calibrate CT scans was equivalent to traditional phantom-based calibration [17]. If assumptions are made about the density of bone and included when deriving phantomless calibration, the results become less accurate as showed by the correlation analysis in the pilot study written by Eggermont et al. [22]. There were several limitations to this study. CT scans of the proximal femur region include a limited choice of tissues to segment: adipose tissue, skeletal muscle, aortic blood, and in some cases the bladder is empty. In addition to population variance across patients, tissues also depend on a variety of specific variables such as: hydration level [20], patient pathologies [43], heterogeneous distributions of muscle and fat [20], i.e. contrast [19,44], exercise habits, and body mass index. The cohort studied here did not include patient details about exercise habits, body mass index or comorbidities. Future studies should consider a more detailed examination of factors known to cause variance across patients and a larger sample size to further develop the phantomless calibration methodology. In this study, GE LightSpeed family CT scanners were used to demonstrate the calibration process. CT scanners from other manufacturers were not analysed due to lack of available data. Future work should consider a multi-centre study comparing the same model of CT scanner across different hospitals and consider CT scanners from other manufacturers. Also of note was the small sample size of available calibration curves for comparison.

This study did not examine several potential confounding variables. When reassigning pre-fracture/control pairings, researchers were not blind to CTXA, a method for measuring areal BMD from CT data mathematically equivalent to dual-energy X-ray absorptiometry, density measurements. Stratification accuracy between pre-fracture and control cases when using phantomless calibration was not examined. Additional confounding factors may have been present such as: other diseases, routine exercise habits, differences in body-mass index/height/weight, comorbidities, or different pathologies. These were not considered due to lack of readily available cohort information. Several of these variables could be considered in a prospective study or in a reanalysis of retrospective data prospectively gathered. A more systemic method of manufacturers were not analysed due to lack of available data. Future work should consider a multi-centre study comparing the same model of CT scanner across different hospitals and consider CT scanners from other manufacturers. Also of note was the small sample size of available calibration curves for comparison.

Overall, results derived from the phantomless calibration slopes were a valid substitute for those derived from the inline calibration. When considering FS, the phantomless calibration resulted in a small 7% increase over inline calibration. For BMD, the phantomless calibration resulted in a small 4% increase over inline calibration. The phantomless calibration slopes were consistently comparable with the range demonstrated by the patient specific Mindways Model 3 BMD phantom calibration slopes, with our best method displaying a small bias of 0.028 ± 0.054 HU/(mg/cm²). The study shows the proposed method for phantomless calibration is valid for FEA studies of retrospective cohorts lacking calibration data. This method can be applied opportunistically to CT scans captured for analyses other than hip fracture. Further examination of the error introduced when the proposed method for phantomless calibration is applied in patient specific FEA derived FS should be conducted.

The segmentation measurements, BMD measurements and femoral strength results created during this study are available by contacting the corresponding author.

CRediT authorship contribution statement

C. Winsor: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing - original draft, Funding acquisition. X. Li: Conceptualization, Methodology, Software, Resources, Writing - review & editing, Supervision, Funding acquisition. M. Qasim: Methodology, Software, Formal analysis, Writing - review & editing. C. Henak: Writing - review & editing, Supervision. P.J. Pickhardt: Investigation, Resources, Writing - review & editing. H. Ploeg: Conceptualization, Writing - review & editing, Funding acquisition. M. Viceconti: Conceptualization, Methodology, Formal analysis, Resources, Writing - review & editing, Supervision, Funding acquisition.

Declaration of competing interest

This study was supported by the Whitaker Foundation. The authors declare that they do not have any financial or personal relationships with other people or organizations that could have inappropriately influenced this study.

Acknowledgements

The authors would like to thank the Whitaker Foundation for financial support of this project, the Institute for International Education for overseeing the Whitaker International Program Fellowship, and the funding from the European Commission H2020 CompBioMed2 project (Grant agreement ID: 823712). This study was funded in part by the EPSRC Frontier Engineering Awards, MultiSim and MultiSim2 projects (Grant reference Numbers: EP/K038777X/1 and EP/S032940/1). Additional thanks go to: Leticia Campello, Ashley Pernsteiner, and Emily Cameron for assistance with repeatability assessments for the phantomless methodology; Bara Kubanova for segmenting patient femurs; Carrie Bartels and Baixuan Yang for assistance capturing CT calibration scans; and Keenan Brown for helpful discussions regarding phantom-based calibration and his experience with GE LightSpeed family CT scanners.

References

Bone 143 (2021) 115759


Software I. Mindways, Qpro tm. 2013.


B. Helgason, E. Perilli, E. Schileo, F. Taddei, S. Brynjolfsson, M. Vicenzi, Mathematical relationships between bone density and mechanical properties: a

11
