The use of routine non density calibrated clinical computed tomography data as a potentially useful screening tool for identifying patients with osteoporosis

Christopher John Burke
Manjiri M. Didolkar
Huiman X. Barnhart
Emily N. Vinson

1 Hospital for Joint Diseases, NYU Langone Medical Center, Department of Radiology, Division of Musculoskeletal Radiology, NY, NY, USA
2 Duke University Medical Center, Department of Radiology, Division of Musculoskeletal Radiology, Durham, NC, USA
3 Duke University, Department of Biostatistics and Bioinformatics, Duke Clinical Research Institute, Durham, NC, USA

Address for correspondence:
Christopher J. Burke
Hospital for Joint Diseases, NYU Langone Medical Center, Department of Radiology, Division of Musculoskeletal Radiology
301 E 17th St, NY, NY 10003, USA
Phone: 212-598-6373 - Fax: 212-598-6125
E-mail: christopher.burke@nyumc.org

Summary

Objectives. To evaluate whether lumbar vertebral body density CT attenuation values measured in Hounsfield Units (HUs) on routine Computed Tomography (CT) examinations can be reliably measured with limited variability, and to evaluate for a correlation between HUs and bone mineral density as measured by dual energy X-ray absorptiometry (DXA) scan.

Methods. Retrospective review of a total of 249 routine MDCT examinations, performed to measure Hounsfield Units (HUs) at the first non-rib bearing lumbar vertebral body on axial images, cross-referenced to the lateral scout image.

Results. The overall ICC and RC for intra-reader variability on CT HU were 0.987 (95% CI 0.973 - 0.999) and 15.664 (95% CI 11.66-19.67). The overall ICC and RDC for inter-reader variability on CT HU were 0.952 (95% CI 0.892 - 0.999) and 30.20 (95% CI 23.73 - 34.48). The ICC and RC for interscanner variability were 0.98 (95% CI 0.95 - 0.99) and 16.57 (95% CI 13.13 - 22.85). The correlation between the L1 HUs and L1 BMD, L1 t-score, and overall t-score was 0.437, 0.392, and 0.400, respectively.

Conclusions. CT attenuation values of the first lumbar vertebra can be measured on routine abdomen CTs with limited variability despite multiple readers and scanners. Correlation between HU and BMD as measured by DXA scan was weakly positive, and by this method measuring the density of a lumbar vertebral body from a routine MDCT scan does not provide the sensitivity or specificity necessary for a screening test. However above a certain measured value (180 HU), patients have a low chance of osteoporosis and therefore may not need additional screening, potentially limiting radiation exposure and cost.

KEY WORDS: computed tomography; bone mineral density; osteoporosis; dual-energy X-ray absorptiometry

Introduction

Bone mineral density (BMD) is inversely proportional to fracture risk and is currently the best predictor of fracture (1-8). Both dual-energy X-ray absorptiometry (DXA) (9) and quantitative CT (10-21) can be used for BMD quantification. Bone mineral content measurement using computed tomography (CT) has been addressed previously (10-26), in particular with regards to quantitative CT techniques (14-18). BMD measurements obtained with CT have been shown in vitro to correlate with fracture patterns and failure loads (14) and biomechanical bone strength can be estimated from quantitative CT (QCT) (19-21). In addition, data obtained from routine, phantomless MDCT examinations has been utilized to extrapolate bone mineral density by comparing to scanner model specific QCT data and using a conversion equation with a model specific conversion factor (27, 28). Post-processing of data from phantomless dual-energy CT studies has also enabled detailed assessment of trabecular bone mineral density in the lumbar spine (29). While these complex processes can be automated to some extent, these methods of determining BMD are not currently widely available. Recently, several studies have emerged proposing a potential correlation between bone mineral density, and density data from routine clinical CT examinations performed for other purposes (23-26), without utilization of a phantom, dual-energy source, or calibration equation. These studies suggest that density measured in Hounsfield units (HU) on routine clinical CT scans is an accurate, reliable, and potentially alternative method of opportunistic screening for osteoporosis, without additional radiation or cost (23-26).

We sought to determine if a simple method of assessing the attenuation characteristics of bone, utilizing the region-of-interest (ROI) tool available on virtually all radiology workstations, obtained from phantomless MDCT exams has been utilized to extrapolate bone mineral density data by comparing to scanner model specific QCT data and using a conversion equation with a model specific conversion factor (27, 28). Post-processing of data from phantomless dual-energy CT studies has also enabled detailed assessment of trabecular bone mineral density in the lumbar spine (29). While these complex processes can be automated to some extent, these methods of determining BMD are not currently widely available. Recently, several studies have emerged proposing a potential correlation between bone mineral density and density data from routine clinical CT examinations performed for other purposes (23-26), without utilization of a phantom, dual-energy source, or calibration equation. These studies suggest that density measured in Hounsfield units (HU) on routine clinical CT scans is an accurate, reliable, and potentially alternative method of opportunistic screening for osteoporosis, without additional radiation or cost (23-26). We sought to determine if a simple method of assessing the attenuation characteristics of bone, utilizing the region-of-interest (ROI) tool available on virtually all radiology workstations, obtained from routine multidetector CT (MDCT) examinations of the abdomen obtained for other clinical purposes, may correlate with BMD as measured by DXA, in the hopes of identifying patients who might benefit from formal BMD assessment and, if appropriate, treatment.

Subjects and methods

Institutional review board approval and waiver of informed consent were obtained for this study. Retrospective review of a total of 249 MDCT examinations, obtained for other clinical indications such as abdominal pain, was performed to measure Hounsfield Units (HU) at the first non-rib bearing lumbar vertebral body on axial images, cross-referenced to the lateral scout image (Figure 1 A and B). Patients with extensive post-operative changes obscurring the first non-rib bearing vertebral body, and patients with lesions or fractures involving the first non-rib bearing vertebral body,
Regions of interest (ROIs) were placed in the trabecular portion of the vertebral body, at approximately the mid-pedicle level. Cortical bone and the basivertebral plexus were avoided. The ROIs ranged in area from 50-100 mm². Scans were performed on any one of several clinical CT scanners available, including GE 16-slice, GE 64-slice, and Siemens 64-slice dual source. There were four parts in this study.

Part one: intra-observer and inter-observer variability
The first part of the study assessed intra-observer and inter-observer variability. Only patients over the age of 50 years were included in this portion of the study. A group of consecutive NECT examinations of the abdomen and pelvis were obtained by querying the PACS archival system by date and examination type. Twenty-six examinations performed in 26 different patients were utilized for this portion of the study. The mean age of the patients was 65.5 years (range, 50 to 81 years). For each examination, HU measurements were made at the mid-pedicle L1 level as described above; measurements were replicated three times by three different radiologist readers, each with different levels of experience (two diagnostic radiology residents with 1 year and 4 years of experience, and one musculoskeletal-trained radiology attending with 10 years of experience), at three different time points. The readers were blinded to the measurements made by other readers, and to their own measurements at other time points.

Part two: inter-scanner variability
The second part of the study assessed inter-scanner variability. Scanning records were queried to identify inpatients who had undergone two separate abdomen NECT examinations on two different MDCT scanners at different time points, with the scans occurring within 15 days of each other (range, 1-15 days). Twenty-six patients were identified meeting these inclusion criteria. For this part of the study, the mean patient age was 48 years (range, 18 to 78 years). One reader measured HUs at the L1 level on each of these studies as described above to assess for inter-scanner variability. The reader was blinded to the results and to previous measurements.

Part three: variability introduced by intravenous contrast enhancement
The third part of this study assessed for variability introduced by intravenous contrast enhancement. Scanner records were queried to identify 26 patients who had undergone a four-phase protocol abdomen MDCT examination. For this part of the study, the mean patient age was 64 years (range, 50 to 78 years). One reader measured HUs at the L1 level, by the method described above, on these 26 abdomen MDCT examinations at each of 4 different phases of IV contrast enhancement: unenhanced phase, hepatic arterial phase, portal venous phase, and delayed phase. The reader was blinded to the results and to previous measurements.

Part four: correlation of HUs and BMD
The fourth part of the study assessed for correlation between the HU measurement obtained at the first non-rib bearing vertebral body on MDCT to the L1 BMD measured on DXA (Hologic, Bedford, MA). Billing records were queried to identify patients who had undergone both abdomen NECT examination and DXA measurement of BMD within 6 months of each other. For this portion of the study, 171 patients were identified; the mean patient age was 71 years (range, 50 to 92 years). One reader measured HUs on the NECT at the L1 level, by the method described above, on each case. BMD and t-score data were subsequently collected from the DXA results via the electronic medical record. The reader was blinded to the DXA data while making the HU measurements.

Statistical analysis
The agreement indices used for assessing intra-observer, inter-observer and inter-phase variability were intra-class correlation coefficient (ICC), repeatability coefficient (RC), reproducibility (RDC) and limits of agreement (LOA). The ICC is an index of relative agreement providing percent of total variability (sum of between-subject variability and random error variability) explained by the between-subject variability. The higher the value of ICC, the better the agreement and thus less random error variability. ICC ≥0.8 is considered as indicating a substantial agreement between measurements and ICC<1.0 indicates perfect agreement. The RC or

Copyright © 2016 CIC Edizioni Internazionali Unauthorized reproduction of this article is prohibited.
RDC is an index of absolute agreement providing the expected difference between any two measurements for 95% of subjects. The lower the value of RC or RDC, the better is the agreement and less random error variability. If RC or RDC is less than or equal to an acceptable/tolerable error between two measurements based on the subject matter, then the agreement is considered to be excellent. If one expects systematic shift such as the measurements in different phases, the LOA index rather than ICC/RC/RDC was used because the asymmetric limits of LOA would indicate this systematic shift. The limits of this index provides an interval where expected differences between any two measurements for 95% of subjects.

For the first part of the study, the ICC and RC were used to assess intra-observer variability by reader at each of the three times. Because RC and ICC are functions of within-subject variability (assuming that the between-subject variability remains the same), testing equality of RCs or equality of ICCs is equivalent to testing equality of within-subject variability. We used generalized estimating equations approach to test the equality of the within-subject variability where the within-subject sample variance for each subject is the response data. If this test is not significant, an overall ICC or RC for intra-observer agreement was computed on the relevant variance components of the three-way ANOVA random effects model where patient, reader and time are the factors for main and all possible interaction terms. The 95% confidence interval (CI) was based on the percentile method from 1000 bootstrap samples taken at the patient level. Similarly, if there were no significant differences between sequence of repetition or time, an overall ICC and RC with 95% CI for inter-observer variability were computed based on the relevant variance components of the same three-way ANOVA random effects model. The ICC and RC were also used to assess inter-reader reliability for different times for 95% of subjects.

For the third part of the study, we expected a possible shift in measurements during different phases, the LOA was then used to assess the expected differences of the HUs measured at any two phases for 95% subjects. For the fourth part of the study, Pearson correlation coefficient was used to assess the correlation between the HU measured on MDCT and the BMD measured on DEXA. A t-score of BMD of -2.5 is used to define osteoporosis and the area under the receiver characteristic curve (AUC) is used to assess the ability of HU measured on MDCT on classifying osteoporosis. All analyses were performed in SAS version 9.2.

Results

Part one: intra-observer and inter-observer variability

The intra-observer reliability for a reader at a time point for HU measurements demonstrated an ICC range of 0.983-0.996 and RC range of 8.77-18.62 (Table 1). There was no statistical significant difference between the ICCs and the RCs by different readers at different times. The overall ICC and RC for intra-observer variability are 0.987 (95% CI 0.973 - 0.999) and 15.66 (95% CI 11.66 - 19.67), respectively, indicating substantial intra-observer reliability.

The inter-reader reliability by time and replication demonstrated an ICC range of 0.938-0.969 and RC range of 24.22-36.44 (Table 2). There was no statistically significant difference between the inter-observer ICCs and the RCs at different times and replications. The overall ICC and RC are 0.952 (95% CI 0.942 - 0.962) and 30.17 (95% CI 26.32 - 34.00), respectively, indicating substantial inter-observer reliability.

Part two: inter-scanner variability

Twenty-five of the 26 patients with NECTs on different CT scanners demonstrated a difference within 13 HUs in the 2 measurements from 2 different CT scanners by the same observer; one patient had a difference of 20 HU. The ICC was 0.98 (95% CI 0.95 - 0.99) and the RC was 16.67 (95% CI 13.13 - 22.85) indicating substantial inter-scanner reliability.

Part three: variability introduced by intravenous contrast enhancement

HU measurements at the 4 different phases of contrast enhancement demonstrated the greatest disagreement in HU measurement between the unenhanced and portal venous phases of contrast, with average mean difference of -24.5 and LOA of -46.8, -2.22 for expected differences of 95% subjects (Table 3).

Part four: correlation of HUs and BMD

The Pearson correlation coefficient between the L1 vertebral body HU from the MDCT examination and the L1 BMD measured by DXA was 0.437. The Pearson correlation coefficient between the L1 vertebral body HU and the L1 t-score was 0.392. The Pearson correlation coefficient between the L1 vertebral body and the overall t-score was 0.400. These values correspond to a weakly

### Table 1 - Intra-observer Variability: replication reliability by reader and time.

<table>
<thead>
<tr>
<th>Reader</th>
<th>Time</th>
<th>ICC</th>
<th>RC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.996</td>
<td>8.77</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>0.983</td>
<td>18.62</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>0.988</td>
<td>14.43</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.989</td>
<td>14.35</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0.984</td>
<td>17.94</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>0.995</td>
<td>9.21</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0.989</td>
<td>14.35</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>0.984</td>
<td>17.94</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>0.987</td>
<td>15.21</td>
</tr>
</tbody>
</table>

### Overall

<table>
<thead>
<tr>
<th>ICC</th>
<th>RC</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.987</td>
<td>15.66</td>
</tr>
</tbody>
</table>

### Table 2 - Inter-observer Variability: inter-reader reliability by time and replication.

<table>
<thead>
<tr>
<th>Time</th>
<th>Replication</th>
<th>ICC</th>
<th>RC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.949</td>
<td>31.23</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>0.943</td>
<td>32.59</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>0.938</td>
<td>36.44</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.968</td>
<td>24.22</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0.949</td>
<td>31.57</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>0.960</td>
<td>27.13</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0.968</td>
<td>24.22</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>0.949</td>
<td>31.57</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>0.942</td>
<td>32.86</td>
</tr>
</tbody>
</table>

### Overall

<table>
<thead>
<tr>
<th>ICC</th>
<th>RC</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.952</td>
<td>30.17</td>
</tr>
</tbody>
</table>

### Table 3 - IV Contrast Enhancement Variability: agreement between any 2 phases.

<table>
<thead>
<tr>
<th>Phase comparison</th>
<th>Mean Difference</th>
<th>Limits of Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phase U vs A</td>
<td>-16.8</td>
<td>-41.6, 8.06</td>
</tr>
<tr>
<td>Phase U vs PV</td>
<td>-24.5</td>
<td>-50.0, -1.22</td>
</tr>
<tr>
<td>Phase U vs D</td>
<td>-10.2</td>
<td>-32.3, 12.0</td>
</tr>
<tr>
<td>Phase A vs PV</td>
<td>-7.77</td>
<td>-27.4, 11.89</td>
</tr>
<tr>
<td>Phase A vs D</td>
<td>6.61</td>
<td>-16.4, 29.61</td>
</tr>
<tr>
<td>Phase PV vs D</td>
<td>14.38</td>
<td>-4.65, 33.70</td>
</tr>
</tbody>
</table>

U - unenhanced, A - hepatic arterial, PV - portal venous, D - delayed.
Recent published studies suggest that MDCT can be used as an opportunistic screening tool with limited variability (23-26). While our results are not surprising when one considers the potential variability in attenuation inherent in clinical CT scanning. The CT number (measured in Hounsfield units) assigned to each pixel of a reconstructed CT image represents the average linear attenuation coefficient of the corresponding volume element imaged (27). Thus, the HU of a vertebral body represents an average of the linear attenuation coefficients of the mineral, collagen, soft tissue, water, and fat in the vertebral body (10, 11). While this CT number is primarily dependent on the atomic number and density of the material imaged and the tube voltage utilized, many scanner and patient factors interact in complex ways to determine attenuation (27).

Several differences in design between our study and these studies may explain the differences in our results. These studies included subjects with spinal trauma or vertebral body compression fractures, while these patients were excluded from our study. Schrieber et al. utilized MDCT where bone reconstruction algorithms were used whereas we utilized routine MDCTs without bone reconstructions. Schrieber et al. also utilized ROIs which were larger and included different areas of the vertebral bodies, while our ROIs were smaller, more random, and only were placed in the mid-pedicle plane. We also did not use mean values of ROIs. Since we demonstrated excellent reliability in the first two parts of our study, we made only one measurement to compare to BMD and t-score from the DXA results, whereas Krapipinger et al. was the fact that they used a calibration device to convert the HU measurement to a BMD value. Our study did not utilize a calibration device.

Our study design suffers from several limitations. Importantly the gold standard against which we were measuring the performance of uncalibrated CT, namely DXA, is itself imperfect with inaccuracies related to body size dependence and degenerative and hypertrophic changes in the targeted elderly population. A comparison with standard QCT may therefore be more meaningful. Furthermore, if the goal is fracture risk prediction rather than specifically WHO-based diagnostic classification, it could be that this simplified CT approach, flawed as it may be, still is a better predictor or discriminator of fracture than the DXA BMD value. There was a selection bias of age of patient for Part 1 and Part 3; the

Discussions

Our data suggests that patients with osteoporosis are more likely to have a measured density that is below 165 HU (or below 180 adjusting for intra-observer variability), with 100% sensitivity. However, many patients with normal bone mineral density by DXA had a measured density from the MDCT below 165 HU, thus using 165 HU as a cut-off number has poor specificity in the detection of osteoporosis with false positive rate of 30%. This is similar to Pickhardt et al. who reported that an L1 CT-attenuation threshold of 160 HU or less was 90% sensitive (and a threshold of 110 HU was more than 90% specific) for distinguishing osteoporosis from osteopenia and normal BMD (25).

Several differences in design between our study and these studies may explain the differences in our results. These studies included subjects with spinal trauma or vertebral body compression fractures, while these patients were excluded from our study. Schrieber et al. utilized MDCT where bone reconstruction algorithms were used whereas we utilized routine MDCTs without bone reconstructions. Schrieber et al. also utilized ROIs which were larger and included different areas of the vertebral bodies, while our ROIs were smaller, more random, and only were placed in the mid-pedicle plane. We also did not use mean values of ROIs. Since we demonstrated excellent reliability in the first two parts of our study, we made only one measurement to compare to BMD and t-score from the DXA results, whereas Schrieber et al. used an average of the lumbar vertebral body HU measurements. One difference to the study by Krapipinger et al. was the fact that they used a calibration device to convert the HU measurement to a BMD value. Our study did not utilize a calibration device.

Our study design suffers from several limitations. Importantly the gold standard against which we were measuring the performance of uncalibrated CT, namely DXA, is itself imperfect with inaccuracies related to body size dependence and degenerative and hypertrophic changes in the targeted elderly population. A comparison with standard QCT may therefore be more meaningful. Furthermore, if the goal is fracture risk prediction rather than specifically WHO-based diagnostic classification, it could be that this simplified CT approach, flawed as it may be, still is a better predictor or discriminator of fracture than the DXA BMD value. There was a selection bias of age of patient for Part 1 and Part 3; the
The use of routine non density calibrated clinical computed tomography data as a potentially useful screening tool for identifying patients with osteoporosis

patients were 50 years or older. The patient’s gender, ethnicity (which the t-score does incorporate) and size were unaccounted for. In addition, while patients with prior lumbar spine surgery, lumbar spine fractures, and spinal metastasis were excluded, we did not attempt to exclude those with metabolic disease processes. When correlating the HUs with BMD from DXA scans, we did not include the cortex of the vertebral body in the HU measurements, while DXA includes cortices in the calculated BMD measurement. We also did not measure the overall HU in the first four lumbar vertebral bodies that DXA scans measure to obtain a measurement. There was a variable placement of ROI in vertebral body; depending on the area where the ROI was placed, more fatty marrow or red marrow may have been selected, which may have changed the HU measurement. Another limitation of this study is its retrospective nature.

A next step may be taking an average of HU measurements of all the lumbar vertebral bodies as measured by DXA, and correlating this with the overall t-score. In addition, including the cortex in the HU measurement as done DXA scans, and adjusting for patient’s size may allow greater accuracy. Measuring the fat content in vertebral bodies on MDCT may also correlate better with BMD measurements and risk of fracture, similar in principle to a study by Tang et al., which correlated an increase in vertebral body marrow fat content measured by magnetic resonance spectroscopy and diffusion weighted magnetic resonance imaging with a decrease in bone density and t-scores on DXA (30). Future work could use dual headed CT to correct for marrow fat (12, 31).

In summary, in contrast to several recent studies, our data suggest the correlation between HU and BMD to be only weakly positive and by this method measuring the density of a lumbar vertebral body from a routine MDCT scan does not provide the sensitivity or specificity necessary for a screening test for osteoporosis. However we did find that above a certain measured density (180 HU), patients have a normal BMD by DXA. Furthermore measurements may be obtained with acceptable and limited variability, despite the use of multiple readers and multiple scanners. Therefore it is possible that, after obtaining a routine MDCT for other clinical purposes, patients with a HU measurement above 180 in the lumbar spine may not need additional screening for osteoporosis, potentially limiting radiation exposure and cost.

Conflict of interest

No conflict of interest.

References


