End-to-End Semi-Supervised Opportunistic Osteoporosis Screening Using Computed Tomography

Jieun Oh¹, Boah Kim², Gyutaek Oh², Yul Hwangbo¹, Jong Chul Ye³

¹Healthcare AI Team, National Cancer Center, Goyang; ²Department of Bio and Brain Engineering, ³Kim Jaechul Graduate School of AI, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea

Background: Osteoporosis is the most common metabolic bone disease and can cause fragility fractures. Despite this, screening utilization rates for osteoporosis remain low among populations at risk. Automated bone mineral density (BMD) estimation using computed tomography (CT) can help bridge this gap and serve as an alternative screening method to dual-energy X-ray absorptiometry (DXA).

Methods: The feasibility of an opportunistic and population agnostic screening method for osteoporosis using abdominal CT scans without bone densitometry phantom-based calibration was investigated in this retrospective study. A total of 268 abdominal CT-DXA pairs and 99 abdominal CT studies without DXA scores were obtained from an oncology specialty clinic in the Republic of Korea. The center axial CT slices from the L1, L2, L3, and L4 lumbar vertebrae were annotated with the CT slice level and spine segmentation labels for each subject. Deep learning models were trained to localize the center axial slice from the CT scan of the torso, segment the vertebral bone, and estimate BMD for the top four lumbar vertebrae.

Results: Automated vertebra-level DXA measurements showed a mean absolute error (MAE) of 0.079, Pearson’s r of 0.852 (P<0.001), and R² of 0.714. Subject-level predictions on the held-out test set had a MAE of 0.066, Pearson’s r of 0.907 (P<0.001), and R² of 0.781.

Conclusion: CT scans collected during routine examinations without bone densitometry calibration can be used to generate DXA BMD predictions.

Keywords: Osteoporosis; Opportunistic screening; Bone mineral density; Dual-energy X-ray absorptiometry; Deep learning

INTRODUCTION

Osteoporosis is a common condition that causes bone fragility. With increasing life expectancy and a graying population across the world, more individuals are projected to be at risk for this condition [1,2]. Osteoporotic fractures are associated with a significantly increased risk of long-term disability [3] and mortality [4,5]. In the United States, approximately 21% of women and...
32% of men aged 65 or older die within 1 year of a hip fracture [6]. Although the condition is both preventable and treatable with early detection as well as its associated socioeconomic burden, osteoporosis remains underdiagnosed [7].

The gold standard modality for estimating bone mineral density (BMD) and diagnosing osteoporosis is dual-energy X-ray absorptiometry (DXA) [8-10]. However, timely diagnosis of this condition is difficult as osteoporosis is asymptomatic [10], and individuals at risk of fracture often have other comorbidities that require more immediate care [11]. To address this disconnection between disease prevalence and diagnosis, previous studies proposed manual and machine learning-based opportunistic osteoporosis screening methods using computed tomography (CT) as an alternative to DXA [12-14]. These methods do not require additional radiation, cost, or, if completely automated, input from health professionals. However, as these systems do not generate end-to-end predictions [15,16] or population agnostic BMD using DXA (BMD_{DXA}) estimates [17-20], there are limitations in applying these results to real-world clinical practice.

This study proposes an automated, end-to-end, deep learning-based method for opportunistic osteoporosis screening of the top four lumbar vertebrae using abdominal CT scans. In contrast to the results of previous studies on CT-based osteoporosis screening, our system generates BMD_{DXA} estimates that are population agnostic and do not rely on calibration using bone densitometry phantoms.

**METHODS**

**Data**

This retrospective study was approved by the Institutional Review Board of the National Cancer Center in the Republic of Korea (NCC2020-0165), and all identifying information was anonymized before analysis. The data used in this study included 367 (198 females and 169 males) contrast-enhanced abdominal CT scans of patients who underwent routine cancer screening between 2010 and 2019 at National Cancer Center in the Republic of Korea. Of these subjects, 268 had associated lumbar DXA values collected within 180 days of their CT scan, irrespective of the imaging order. Patients with spinal implants or DXA z-scores greater than 3.3 or less than –3.3 were excluded.

**Table 1. Demographic Characteristics**

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>All subjects</th>
<th>Male</th>
<th>Female</th>
<th>Train</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subjects</td>
<td>268</td>
<td>122</td>
<td>146</td>
<td>170</td>
<td>98</td>
</tr>
<tr>
<td>Age, yr</td>
<td>58.86±12.56</td>
<td>63.05±11.89</td>
<td>55.36±12.01</td>
<td>58.22±13.56</td>
<td>59.97±10.51</td>
</tr>
<tr>
<td>Height, cm</td>
<td>161.35±8.54</td>
<td>167.07±7.08</td>
<td>156.57±6.47</td>
<td>160.96±8.53</td>
<td>162.02±8.51</td>
</tr>
<tr>
<td>Weight, kg</td>
<td>61.60±11.53</td>
<td>66.15±12.37</td>
<td>57.79±9.19</td>
<td>60.47±11.17</td>
<td>63.56±11.88</td>
</tr>
<tr>
<td>BMI status</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Underweight</td>
<td>12 (4.48)</td>
<td>7 (5.74)</td>
<td>5 (3.42)</td>
<td>10 (5.88)</td>
<td>2 (2.04)</td>
</tr>
<tr>
<td>Normal</td>
<td>116 (43.28)</td>
<td>42 (34.43)</td>
<td>74 (50.68)</td>
<td>75 (44.12)</td>
<td>41 (41.84)</td>
</tr>
<tr>
<td>Overweight</td>
<td>33 (12.31)</td>
<td>12 (9.84)</td>
<td>21 (14.38)</td>
<td>17 (10.0)</td>
<td>16 (16.33)</td>
</tr>
<tr>
<td>Obese</td>
<td>107 (39.93)</td>
<td>61 (50.0)</td>
<td>46 (31.51)</td>
<td>68 (40.0)</td>
<td>39 (39.80)</td>
</tr>
<tr>
<td>BMD values</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L1–L4 BMD, g/cm²</td>
<td>0.93±0.16</td>
<td>0.99±0.16</td>
<td>0.88±0.15</td>
<td>0.92±0.15</td>
<td>0.94±0.19</td>
</tr>
<tr>
<td>L1 BMD, g/cm²</td>
<td>0.85±0.15</td>
<td>0.91±0.15</td>
<td>0.81±0.14</td>
<td>0.85±0.14</td>
<td>0.85±0.17</td>
</tr>
<tr>
<td>L2 BMD, g/cm²</td>
<td>0.91±0.17</td>
<td>0.97±0.17</td>
<td>0.86±0.15</td>
<td>0.90±0.15</td>
<td>0.92±0.20</td>
</tr>
<tr>
<td>L3 BMD, g/cm²</td>
<td>0.95±0.17</td>
<td>1.01±0.17</td>
<td>0.90±0.16</td>
<td>0.94±0.16</td>
<td>0.96±0.19</td>
</tr>
<tr>
<td>L4 BMD, g/cm²</td>
<td>0.98±0.19</td>
<td>1.05±0.19</td>
<td>0.92±0.16</td>
<td>0.97±0.17</td>
<td>1.00±0.21</td>
</tr>
<tr>
<td>BMD class</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normal</td>
<td>151 (56.34)</td>
<td>84 (68.85)</td>
<td>67 (45.89)</td>
<td>99 (58.24)</td>
<td>52 (53.06)</td>
</tr>
<tr>
<td>Osteopenia</td>
<td>88 (32.84)</td>
<td>32 (26.23)</td>
<td>56 (38.36)</td>
<td>53 (31.18)</td>
<td>35 (35.71)</td>
</tr>
<tr>
<td>Osteoporosis</td>
<td>29 (10.82)</td>
<td>6 (4.92)</td>
<td>23 (15.75)</td>
<td>18 (10.59)</td>
<td>11 (11.22)</td>
</tr>
</tbody>
</table>

Values are expressed as mean± standard deviation or number (%).
BMI, body mass index; BMD, bone mineral density.
from this study. We did not include cases with a suspected fracture on CT scout films or a suspected fracture on DXA with a T-score difference of more than 1 between adjacent vertebrae.

Table 1 contains additional demographic information on the DXA subject population. Subjects were grouped into normal, osteopenia, and osteoporosis groups using their L1–L4 DXA T-score values and the World Health Organization (WHO) diagnostic criteria. Body mass index (BMI) status was based on the WHO BMI cutoffs for Asian populations, where a value below or equal to 18.5 is underweight, between 18.5 and 23 is normal, between 23 and 27.5 is overweight, and greater than 27.5 is obese [21].

The CTs and Convolution Kernels used were: (1) Discovery CT 750HD with Convolution Kernel: Standard (GE Healthcare, Waukesha, WI, USA) (n=238); (2) Brilliance with Convolution Kernel: B (Philips Medical Systems, Best, the Netherlands) (n=109); (3) SOMATOM Definition Edge with Convolution Kernel: B30f (Siemens Healthcare, Forchheim, Germany) (n=20).

All CTs were scanned 80 seconds after injection of Omnipaque (Iohexol, GE Healthcare) 300 contrast medium (Venous phase). Peak tube voltage was 120 kVp (n=347) or 100 kVp (n=20). The tube voltage used for all abdominal CT images was 120 kVp and the tube current (milliamperere) was adjusted automatically by the thickness of the patient’s body for radiation dose optimization.

Lumbar DXA scans were performed using QDR 4,500 W (n=165), Horizon W (n=98), and Discovery W (n=5) scanners manufactured by Hologic (Marlborough, MA, USA). For each subject, slice number annotations and vertebral segmentation mask labels were assigned to the central axial CT slice of the L1, L2, L3, and L4 vertebrae.

**Experimental setup**

Fig. 1 illustrates the CT-based opportunistic osteoporosis screening method. The system generated end-to-end BMD using convolutional neural network (BMDcNN) estimates from abdominal...
CT scans via three successive steps: localization, segmentation, and estimation. The localization subtask takes the abdominal CT scan of a subject as the input and outputs the center axial CT slice locations for the L1, L2, L3, and L4 vertebrae as integer values. For each of the four predicted center slices, the system takes one slice above and two slices below the detected center, segmenting the vertebral bone into four slices during the segmentation step. The segmented slices corresponding to each lumbar vertebra were cropped, concatenated counterclockwise, and used as the input for the estimation step to generate the BM-DCNN predictions.

**Lumbar spine localization**

The localization subtask uses deep learning-based regressors to detect the center axial slice locations of the L1, L2, L3, and L4 spines. Our localization method is a two-step process that utilizes both frontal and sagittal maximum intensity projections (MIPs) and two regression models. In contrast to the previous work on CT-based lumbar spine localization by Belharbi et al. [22] and Kanavati et al. [23], our method outputs vertebral-level predictions for the top four lumbar vertebrae rather than just the center L3 slice location. Fig. 2 shows the overall flow of the proposed method.

The two stages of the localization method are trained separately and combined during the inference phase. In the first stage, a neural network outputs the top axial slice location of the L1 vertebra using frontal MIP, and the second model predicts the center slice locations of the L1, L2, L3, and L4 vertebrae during the second stage. The frontal MIP used in the first stage was center cropped along the x-axis and zero padded along the y-axis to generate a 160×256 pixel image. In the second stage, the sagittal MIP was cropped from the 120th pixel to the 280th pixel along the x-axis and 80 slices from the top L1 slice along the y-axis to generate a 160×80-pixel region of interest. During training, the sagittal MIP was cropped, starting from a random slice between 5 and 15 above the top L1 location ground-truth, and the MIP was cropped from 10 slices above the top L1 slice prediction in the first step of the localization subtask. Both stages use VGG19-based regressors [24], with a channel attention block applied before each pooling layer. Additional implementation details are outlined in Supplemental Fig. S1.

**Vertebral segmentation**

Spine BMD_{DXA} measurements are known to be affected by body composition [25]; therefore, our system segments the vertebra for each axial CT slice using a U-Net-based model [26], such that the subsequent estimation subtask will only use bone images as input. However, there is a trade-off between improving model performance with more labeled samples and the resource demands associated with manual data annotation. To address this disconnection, we take a semi-supervised approach for this subtask, where we only use one segmentation mask for each
Fig. 3. Training data for semi-supervised vertebra segmentation. For all subjects, the central axial computed tomography slice was annotated with segmentation slice labels. All axial slices for a given vertebra were used regardless of segmentation ground-truth availability during training.

Fig. 4. Data generation process for the bone mineral density estimation subtask. (A) Four computed tomography (CT) slices were selected from the detection subtask and their corresponding masks were generated from the segmentation subtask. (B) A 196×196 patch centered on the center of mass of the binary segmentation mask is cropped from each CT slice. (C) Cropped CT slices are concatenated counterclockwise to form a 392×392 input image.
vertebra but use all available axial slices during training via the Mumford-Shah (MS) loss [27]. As shown in Fig. 3, a semi-supervised learning approach allows us to use all the CT slices in a vertebra, leading to an approximately ten-fold increase in training data as compared to a strongly supervised approach using only the center vertebral slices. The loss was calculated using a weighted sum of cross-entropy (CE) and MS losses, where the CE loss was only applied to samples with ground-truth segmentation masks, whereas the MS loss was applied to all training samples. Additional details on the model architecture, loss function, and training procedure are provided in Supplemental Fig. S2.

Bone mineral density estimation

In this study, BMD estimation was considered as an image regression task, where the system generated continuous BMD$_{\text{CNN}}$ predictions for the L1, L2, L3, and L4 vertebrae independently using a DenseNet169-based regressor [28]. Fig. 4 illustrates the image generation process for the BMD estimation subtask. For training, all axial CT slices from the entire range of a given vertebra were segmented, whereas the images used for testing were based on the output predictions from the detection and segmentation subtasks. For a given vertebra, four consecutive axial CT slices were selected to generate one input sample. During training and validation, a sliding window was applied to the slices, such that for a vertebra with $n$ ordered slices, $n-3$ samples containing four adjacent CT slices were generated. In contrast to using only the slices near the center, this approach uses data from the entire vertebra during training while ensuring that slices closer to the start and end of the vertebra, which contain a higher proportion of denser cortical bone, are sampled less frequently. During testing, the center slice prediction for a given vertebra was used from the detection step, and one slice above and two slices below the predicted slice were selected. This choice, in contrast to two slices above and one slice below the detected center, was arbitrary because the differences between the two methods were negligible during model selection. Additional training and architecture details for the estimation subtask are available in Supplemental Methods.

Statistical analysis

The models in this study were evaluated using Dice scores, regression metrics (mean absolute error [MAE], root mean squared error, and mean absolute percentage error), Pearson correlation coefficients, and coefficients of determination. Dice scores were calculated using PyTorch (https://pytorch.org) [29]. Regression metrics and coefficients of determination were evaluated using Python and the scikit-learn [30] package. Pearson $r$ coefficients and their respective $P$ values were acquired using Python and the SciPy [31] package.

RESULTS

Lumbar spine localization

Table 2 presents the results of the proposed localization method at the vertebral and subject levels. Localization performance was evaluated in CT slices with an absolute error. The performance with the channel attention block was listed in Supplemental Table S1.

Vertebral segmentation

For each CT slice, segmentation maps were generated using the largest connected component of the predicted output. Predictions were evaluated using Dice scores. For a segmentation map $A$ and ground-truth label $B$, the Dice score is defined as:

$$\text{Dice}(A, B) = \frac{2|A \cap B|}{|A| + |B|}$$

(1)

Table 2. Spine Localization Results on the Held-out Test Set ($n=98$)

<table>
<thead>
<tr>
<th>Location</th>
<th>Mean</th>
<th>SD</th>
<th>Median</th>
<th>Max</th>
<th>$\geq 10^a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1</td>
<td>0.98</td>
<td>2.26</td>
<td>0</td>
<td>13</td>
<td>4</td>
</tr>
<tr>
<td>L2</td>
<td>1.02</td>
<td>2.26</td>
<td>0</td>
<td>13</td>
<td>4</td>
</tr>
<tr>
<td>L3</td>
<td>1.08</td>
<td>2.38</td>
<td>0.5</td>
<td>14</td>
<td>4</td>
</tr>
<tr>
<td>L4</td>
<td>1.18</td>
<td>2.37</td>
<td>1</td>
<td>13</td>
<td>4</td>
</tr>
<tr>
<td>Total</td>
<td>1.07</td>
<td>2.31</td>
<td>1</td>
<td>14</td>
<td>16</td>
</tr>
</tbody>
</table>

The mean, standard deviation, median, and max columns are statistics on the absolute error in the slices. SD, standard deviation. $^a$The number of predictions with an error greater than or equal to 10 slices.

Table 3. Results of Segmentation Method Evaluated on Vertebral- and Subject-Levels

<table>
<thead>
<tr>
<th>Method</th>
<th>Dice by vertebra</th>
<th>Dice by subject</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.9673</td>
<td>0.9677</td>
</tr>
<tr>
<td>Proposed</td>
<td>0.9683</td>
<td>0.9685</td>
</tr>
</tbody>
</table>

The “baseline” model is a strongly supervised standard U-Net while the “proposed” model is our semi-supervised segmentation model using Mumford-Shah loss.
where $|A \cap B|$ is the area of the overlapping region between $A$ and $B$ and $|A| + |B|$ is the sum of the areas $A$ and $B$. Since the proposed screening method is performed end-to-end, segmentation performance is contingent on the results of the localization subtask (Supplemental Fig. S3). However, the vertebral segmentation masks were evaluated and the Dice scores on the manually labeled center vertebral slices were computed. Table 3 reports the results of our semi-supervised segmentation method at the subject and vertebral levels compared with a standard, strongly supervised U-Net. The detailed results were presented in Supplemental Table S2.

**Bone mineral density estimation**

Table 4 lists the end-to-end performance of the estimation subtask. The 98 subjects in the test set (52 females and 46 males) were represented by 392 image samples generated from 1,568 CT slices, where each image represented one lumbar vertebra. Vertebral-level results were obtained by evaluating the end-to-end BMD$_{\text{CNN}}$ estimates against the BMD$_{\text{DXA}}$ ground truths, while the subject-level results were generated by taking the arithmetic mean of the L1, L2, L3, and L4 BMD$_{\text{CNN}}$ predictions and comparing them to the total lumbar BMD$_{\text{DXA}}$ values. Fig. 5 illustrates these results using regression and Bland-Altman [32] plots.

**DISCUSSION**

This study proposes a deep learning-based system for end-to-end opportunistic osteoporosis screening using abdominal CT scans. Our system detects the center L1, L2, L3, and L4 axial CT slices with a MAE of $1.07 \pm 2.31$ slices, segments the vertebra with a Dice score of 0.968, and generates vertebral-level BMD$_{\text{DXA}}$ estimates with a MAE of 0.079, Pearson $r$ of 0.852 ($P<0.001$), and $R^2$ of 0.714 on the held-out test set. For subject-level predictions, the system output predictions had an MAE of 0.066, Pearson $r$ of 0.907 ($P<0.001$), and $R^2$ of 0.781.

Previous studies have investigated deep learning-based methods for CT-based BMD estimation. Some of these studies relied on extracting features on manually selected CT slices rather than entire studies. Tang et al. [16] proposed a method of screening for osteoporosis using a multiclass classification network to discriminate between normal bone mass, low bone mass, and osteoporosis from segmented axial CT slices, while Yasaka et al. [15] used manually cropped mid-vertebral axial CT slices with circular region-of-interest markers to train a convolutional neural network (CNN) to output continuous BMD$_{\text{DXA}}$ predictions. Other more recent studies took end-to-end approaches to BMD estimation. Pickhardt et al. [33] and Fang et al. [18] presented end-to-end automatic methods to predict quantitative CT (QCT) values, while Liu et al. [19] correlated the mean Hounsfield unit values in regions of interest from low-dose chest CT scans with BMD$_{\text{DXA}}$ values for the thoracic and first two lumbar vertebrae. Krishnaraj et al. [20] used a cascade of two segmentation networks and machine learning-based regression to predict the DXA t-scores for the top four lumbar vertebrae from extracted images. The segmentation performance of our method was compared to other vertebrae with ground-truth lumbar BMD$_{\text{DXA}}$ values. Subject-level results were generated by comparing total BMD$_{\text{DXA}}$ values with averaged L1, L2, L3, and L4 BMD$_{\text{CNN}}$ predictions for a given subject. All Pearson correlation coefficients had $P$ values less than 0.001.

### Table 4. Results of End-to-End BMD$_{\text{DXA}}$ Estimation across All Test Subjects (52 Females, 46 Males)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Pearson $r$</th>
<th>$R^2$</th>
<th>MAE</th>
<th>RMSE</th>
<th>MAPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vertebral-level</td>
<td>$8.522 \times 10^{-1}$</td>
<td>$7.141 \times 10^{-1}$</td>
<td>$7.928 \times 10^{-2}$</td>
<td>$1.076 \times 10^{-1}$</td>
<td>8.578</td>
</tr>
<tr>
<td>Subject-level</td>
<td>$9.066 \times 10^{-1}$</td>
<td>$7.810 \times 10^{-1}$</td>
<td>$6.559 \times 10^{-2}$</td>
<td>$8.662 \times 10^{-2}$</td>
<td>7.087</td>
</tr>
<tr>
<td>L1</td>
<td>$8.637 \times 10^{-1}$</td>
<td>$7.333 \times 10^{-1}$</td>
<td>$6.682 \times 10^{-2}$</td>
<td>$8.568 \times 10^{-2}$</td>
<td>8.149</td>
</tr>
<tr>
<td>L2</td>
<td>$8.918 \times 10^{-1}$</td>
<td>$7.762 \times 10^{-1}$</td>
<td>$7.356 \times 10^{-2}$</td>
<td>$9.458 \times 10^{-2}$</td>
<td>8.056</td>
</tr>
<tr>
<td>L3</td>
<td>$8.697 \times 10^{-1}$</td>
<td>$7.169 \times 10^{-1}$</td>
<td>$8.007 \times 10^{-2}$</td>
<td>$1.027 \times 10^{-1}$</td>
<td>8.396</td>
</tr>
<tr>
<td>L4</td>
<td>$7.734 \times 10^{-1}$</td>
<td>$5.724 \times 10^{-1}$</td>
<td>$9.667 \times 10^{-2}$</td>
<td>$1.396 \times 10^{-1}$</td>
<td>9.711</td>
</tr>
<tr>
<td>Female by vertebra</td>
<td>$8.820 \times 10^{-1}$</td>
<td>$7.599 \times 10^{-1}$</td>
<td>$6.875 \times 10^{-2}$</td>
<td>$8.960 \times 10^{-2}$</td>
<td>8.559</td>
</tr>
<tr>
<td>Female by subject</td>
<td>$9.172 \times 10^{-1}$</td>
<td>$8.055 \times 10^{-1}$</td>
<td>$5.835 \times 10^{-2}$</td>
<td>$7.406 \times 10^{-2}$</td>
<td>7.129</td>
</tr>
<tr>
<td>Male by vertebra</td>
<td>$7.966 \times 10^{-1}$</td>
<td>$5.603 \times 10^{-1}$</td>
<td>$9.118 \times 10^{-2}$</td>
<td>$1.249 \times 10^{-1}$</td>
<td>8.599</td>
</tr>
<tr>
<td>Male by subject</td>
<td>$8.659 \times 10^{-1}$</td>
<td>$6.512 \times 10^{-1}$</td>
<td>$7.379 \times 10^{-2}$</td>
<td>$9.892 \times 10^{-2}$</td>
<td>7.040</td>
</tr>
</tbody>
</table>

Vertebral-level results are based on end-to-end bone mineral density using convolutional neural network (BMD$_{\text{CNN}}$) predictions evaluated independently from other vertebrae with ground-truth lumbar BMD$_{\text{DXA}}$ values. Subject-level results were generated by comparing total BMD$_{\text{DXA}}$ values with averaged L1, L2, L3, and L4 BMD$_{\text{CNN}}$ predictions for a given subject. All Pearson correlation coefficients had $P$ values less than 0.001. BMD$_{\text{DXA}}$, bone mineral density using dual-energy X-ray absorptiometry; MAE, mean absolute error; RMSE, root mean squared error; MAPE, mean absolute percentage error.
voxel volumes. With the exception of Liu et al. [34] and Pickhardt et al. [33], these prior studies relied on strongly supervised methods (i.e., each training sample has a corresponding ground-truth label) for slice detection and region-of-interest extraction tasks. To reduce the cost associated with manual annotation, our work adopts a semi-supervised approach to segmentation, where only the ground-truth masks for the center axial CT slice of each vertebra are provided. We note that while the localization and segmentation approaches used by Liu et al. [34] and Pickhardt et al. [33] require no masks, the respective methods require ana-

![Fig. 5](https://www.e-enm.org) Regression (left) and Bland-Altman (right) plots for bone mineral density using dual-energy X-ray absorptiometry (BMD_\text{DXA}) ground truths and end-to-end bone mineral density using convolutional neural network (BMD_\text{CNN}) predictions. For both regression plots, the 95% confidence interval and 95% prediction interval are represented by the orange dotted line and grey dashed line respectively. (A) Vertebral-level results for L1, L2, L3, and L4 predictions evaluated independently (n=392). (B) Subject-level results for averaged L1 to L4 BMD_{\text{CNN}} predictions against total lumbar BMD_{\text{DXA}} (n=98). SD, standard deviation.
tomical landmarks (i.e., clavicles) that are not present in our data-
set of abdominal CT scans [34] or the use of the lowest ribs [35],
which can be difficult owing to lumbar or absent ribs.

Our system generates end-to-end predictions of $\text{BMD}_{\text{DXA}}$ values from abdominal CT scans rather than DXA t-scores [20], normal/osteopenia/osteoporosis classes [16], or QCT BMD values [17,18]. DXA t-scores and DXA-based osteoporosis classifications are population-dependent, and thresholds for osteopenia and osteoporosis do not always generalize across different groups. Although QCT has certain benefits over DXA, including capturing volumetric rather than areal BMD, the screening modality has lower utilization than DXA and often requires the use of a BMD calibration phantom at the time of the scan. This limits the number of available QCT samples relative to that of DXA, leaving fewer CT-BMD pairs for future training and validation.

This study has a few limitations of its own. All DXA and CT studies were sourced via routine cancer screening at a single oncology specialty center in the Republic of Korea, limiting the scope of our analysis as BMD varies between national and ethnic groups [2,36]. To generate vertebral-level annotations frontal MIP images were compared to lumbar DXA images. This was because of the difficulty introduced by the transitional vertebrae, lumbar ribs, and anatomical markers not being present in abdominal CTs, which would enable annotation with CT images alone. In particular, lumbosacral transitional vertebrae (LSTV)—common anatomical variants of the lumbar spine caused by the lumbarization of the S1 vertebra (leading to the appearance of six lumbar vertebrae) and sacralization of the L5 vertebra (leading to the appearance of four lumbar vertebrae)—make vertebral-level annotation difficult even for trained clinicians [37]. In practice, LSTVs cause errors in imaging [38] and wrong-site surgery [39]. To account for these variations, DXA images were used as ground-truth labels to reflect real-world clinical decision-making. Additionally, we were unable to collect information on family history, menopausal status, physical activity, or other risk factors.

This study describes an end-to-end deep learning-based system to measure BMD using abdominal CT. Furthermore, this study demonstrated that CT scans collected during routine cancer screening without bone densitometry phantoms could generate $\text{BMD}_{\text{DXA}}$ predictions without additional tests.

In this study, we have shown that machine learning can be used to predict BMD at each level of the lumbar spine from routine CT images. Next, we would like to develop a machine learning model that predicts the T-score and $z$-score of L1–L4 from CT images and develop software to diagnose osteoporosis. We also want to develop artificial intelligence software that uses clinical data from electronic health records and CT images to predict the long-term risk of fractures.

**CONFLICTS OF INTEREST**

No potential conflict of interest relevant to this article was reported.

**ACKNOWLEDGMENTS**

This work was funded by Korea Advanced Institute of Science and Technology (KAIST) R&D Program (KI Meta-Conver-
gence Program) 2020 (N10200012) through KAIST and by a grant from the National Cancer Center (2010080 and 2310840) in the Republic of Korea. Boah Kim is now at National Institutes of Health, Bethesda, MD, USA.

**AUTHOR CONTRIBUTIONS**

Conception or design: Y.H., J.C.Y. Acquisition, analysis, or interpretation of data: J.O., B.K., G.O., Y.H., J.C.Y. Drafting the work or revising: J.O., B.K., G.O. Final approval of the manuscript: J.O., B.K., G.O., Y.H., J.C.Y.

**ORCID**

Jieun Oh https://orcid.org/0000-0002-1953-9845
Yul Hwangbo https://orcid.org/0000-0001-7129-2133
Jong Chul Ye https://orcid.org/0000-0001-9763-9609

**REFERENCES**

Osteoporosis Screening Using Computed Tomography


SUPPLEMENTAL METHODS

Introduction

Our system generates end-to-end bone mineral density (BMD) estimates from abdominal computed tomography (CT) scans via three successive steps:

1. Localization: Given a subject’s abdominal CT scan, the system detects the center transverse CT slice from the L1, L2, L3, and L4 vertebrae.
2. Segmentation: For each lumbar vertebra, four CT slices—the center slice, the slice above the center, and two slices below the center—were selected based on the output from the localization step and segmented to retain only the vertebral bone.
3. Estimation: Four cropped regions corresponding to each vertebra are concatenated counterclockwise to generate a BMD using convolutional neural network (DXACNN) prediction for each vertebra independently.

Localization of lumbar spines

Inspired by the work of Belharbi et al. [1] and Kanavati et al. [2], we proposed a method for localizing the center L1, L2, L3, and L4 axial CT slices using maximum intensity projections (MIPs). The MIPs were images generated by taking the highest intensity value along a specific axis for each pixel of the projection. As bones have higher Hounsfield unit (HU) values than soft tissues, MIPs from CT stacks can generate two-dimensional (2D) representations of the bone structure of a subject. In this study, we clipped the intensity values between –100 and 1,000 HU to project both coronal and sagittal MIPs.

Our system takes a two-stage approach for this subtask: the first stage finds the slice corresponding to the top of the L1 vertebra, whereas the second stage outputs the central slice locations of the top four lumbar vertebrae based on the results of the first stage. Both stages were trained separately and combined during inference. Channel attention (CA) was employed and pre-trained networks were not used for slice localization.

Training details

1) Network architecture

Supplemental Fig. S1A illustrates the network architectures of the two models used in the subtask. Our networks are based on the Visual Geometry Group 19 (VGG19) architecture [3], which consists of 3×3 convolution layers, batch normalization [4], rectified linear unit (ReLU) activations, max-pooling layers for feature extraction, and a fully connected layer to generate outputs. In contrast to standard VGG19 models, CA blocks were applied before each pooling layer to improve model performance. Supplemental Fig. S1B shows the architecture of a CA block. First, the global average pooling was applied to the input features. Subsequently, the weight of each channel was calculated as a value between 0 and 1 after the 1×1 convolution and sigmoid layers. The calculated weights were then multiplied by the input feature.

2) Implementation

Both the sagittal and frontal MIPs were normalized to a range of 0 to 1. Horizontal flipping and shifting data augmentations were applied. An Adam optimizer with a momentum $\beta_1=0.9$, $\beta_2=0.999$, mean absolute error (MAE) loss, and a batch size of four were used. The initial learning rate (LR) is set to $1 \times 10^{-4}$ and reduced by half for every 200 epochs. In addition, the neural networks in the first and second stages were trained for 1,000 epochs. Both networks were implemented using PyTorch (https://pytorch.org).

Ablation study for the channel attention

Supplemental Table S1 presents the results of the models trained with and without CA on our test set. The model trained with CA had lower MAE means and standard deviations, as well as fewer outliers, than the models trained without attention.

Lumbar spine segmentation

To reduce the cost associated with annotating a vast number of 2D CT slices and improve the segmentation performance, as shown in Supplemental Fig. S2A, a semi-supervised segmentation method was applied. Since the Mumford-Shah loss [5] does not require ground-truth masks, the network uses many unlabeled CT slices.

Specifically, given an input image $x$ and predicted segmentation map $y$, the network is trained by minimizing the following loss function:

$$\text{Loss}=\alpha L_{CE}+\beta L_{MS}$$

where $\alpha$ and $\beta$ are hyper-parameters, $L_{CE}$ is the cross-entropy loss, and $L_{MS}$ is the Mumford-Shah loss. The hyperparameter $\alpha$ is 1 if the input contains labeled masks and $\alpha=0$ otherwise. Accordingly, as illustrated in Supplemental Fig. S2A, the $L_{CE}$ that requires ground-truth labels and predicted segmentation maps is computed when the networks use annotated data, whereas the Mumford-Shah loss, which requires inputs and network outputs, is applied to all training data. The $L_{CE}$ is computed as

$$L_{CE}(y,g) = -\frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{M} g_{n,m} \log y_{n,m}$$

where $g$ is the ground truth label, $M$ is the number of classes,
and \( N \) is the number of pixels. In addition, under the assumption that the segmentation map \( y \) is the output of the softmax layer of the network, the Mumford-Shah loss can be expressed as:

\[
L_{MS}(x, y) = \sum_{m=1}^{M} \int_{\Omega} |x(n) - c_m| y_m(n) \, dn + \lambda \sum_{m=1}^{M} \int_{\Omega} |\nabla y_m(n)| \, dn \tag{3}
\]

where

\[
c_m = \frac{\int_{\Omega} x(n) y_m(n) \, dn}{\int_{\Omega} y_m(n) \, dn} \tag{4}
\]

This is the average pixel value of the \( m \)-th class of the input images. Therefore, the network was trained in a semi-supervised manner using the proposed loss function (1).

**Training details**

1) **Network architecture**

Supplemental Fig. S2B illustrates the U-Net-like architecture used in the segmentation network. U-Net consists of encoder and decoder paths with skip connections between them and is widely used in image segmentation. The encoder was composed of four repeated convolution blocks for down-sampling. The convolution block has two-unit blocks formed by a series of \( 3 \times 3 \) convolutions: batch normalization [4] and ReLU activation. The decoder was configured similarly to the encoder, with the convolution module containing a convolution transpose layer with a stride of two for up-sampling, which was repeated four times. For each up-sample, the number of channels was halved. The skip connections link the features from the encoder to the up-sampled features by concatenation. In the last layer of the network, a convolution layer with \( 1 \times 1 \) kernels and a softmax layer was applied to generate probability maps for segmentation.

2) **Implementation**

The proposed semi-supervised vertebral segmentation method was implemented in Python using the PyTorch library. We down-sampled the 2D CT images from 512×512 to 256×256 by sub-sampling and augmented the data using horizontal or vertical flipping and \( 90^\circ \) rotations. The network was trained using an Adam optimizer [6] with an initial LR of \( 1 \times 10^{-5} \). The LR was decreased by half after every 20 epochs. The batch size was set to four and the hyperparameter \( \beta \) to 0.01. The parameters of the network were initialized by the weights of a pre-trained network using only labeled data, and the network was trained for 100 epochs using a single NVIDIA Quadro RTX 6000 GPU (NVIDIA, Santa Clara, CA, USA).

**Additional experimental results**

Supplemental Table S2 presents the cross-validation results. When comparing the proposed semi-supervised method to the baseline supervised method, the Dice scores of the proposed method across all validation sets were higher than those of the comparative method. In addition, the average global Dice score of the validation sets using the proposed method showed a 0.12% gain over the baseline.

1) **Qualitative segmentation results**

Supplemental Fig. S3 illustrates the qualitative evaluation results for vertebral segmentation. For both male and female subjects, all lumbar vertebral regions were segmented more accurately using the proposed method than with the supervised method. Comparing the Dice score results of the test dataset for the two models, the proposed semi-supervised method outperformed the baseline.

**Bone mineral density estimation**

The inputs for the estimation task are based on the results of the two preceding steps. Four central axial CT slices were selected from each of the top four lumbar vertebrae and binary vertebral segmentation masks were acquired for the 16 images from the segmentation subtask. The Hadamard product of the mask and CT slice were taken, the pixel intensity values were clipped to \((0,1000) \) HU, and a \( 196 \times 196 \)-pixel area in the image centered on the center of mass of the segmentation mask was cropped. These four images were then concatenated clockwise to generate a \( 392 \times 392 \)-pixel image that was resized to \( 98 \times 98 \)-pixels and min-max normalized by the sample to \((0,1)\).

**Network architecture**

The regression network used for BMD estimation was based on the DenseNet169 architecture [7], with the output layer replaced by four blocks of successive fully connected layers with ReLU activation and batch normalization [4]. The fully-connected layers (from bottom to top) had 1,024, 512, 128, and 8 nodes, respectively. The final output layer is a fully connected layer with a single node and linear activation. All layers in the model were initialized using the Glorot uniform initialization [8].

**Implementation**

The regressor was trained using CT-dual-energy X-ray absorptiometry pairs from 158 subjects and validated using data from 15 subjects. In line with other studies on deep learning in medical imaging [9,10], 30 regressors with a fixed random seed, archi-
Architecture, dataset splits, and training procedure were trained and the five models with the lowest validation loss were assembled. The sliding window method generated 4,727 training and 417 validation samples. During the training, rotation (between \(-10^\circ\) and \(10^\circ\)), horizontal and vertical shifts (by a factor of 0.05), horizontal flipping, and zooming (between 0.95 and 1.05) augmentations were randomly applied. The regressor models were trained using MAE loss with a batch size of 32 and an Adam optimizer \([6]\) with an initial LR of 0.01, a \(\beta_1\) of 0.9, a \(\beta_2\) of 0.999, and an \(\epsilon\) of \(10^{-7}\) for 200 epochs. Both learning-rate scheduling and early stopping were applied based on the validation loss. LR was reduced by a factor of 0.33 if the validation loss did not improve for five epochs, when the LR was greater than 0.0001, and training was stopped if the validation loss did not improve for 15 epochs. The model was trained using Python 3.6, TensorFlow 1.13 (https://www.tensorflow.org/?hl=ko), and CUDA 10.0 (https://developer.nvidia.com/cuda-10.0-download-archive) on an Ubuntu 18.04 (https://releases.ubuntu.com/18.04/) workstation with an Intel i9-10940X processor (https://www.intel.co.kr/), 128 GB of memory, and an NVIDIA RTX Titan GPU.

SUPPLEMENTAL REFERENCES

**Supplemental Table S1.** Results of Spine Localization on 98 Test Data

<table>
<thead>
<tr>
<th>Method</th>
<th>Spine</th>
<th>Mean</th>
<th>SD</th>
<th>Median</th>
<th>Max</th>
<th>≥10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed (without CA)</td>
<td>L1</td>
<td>1.29</td>
<td>2.97</td>
<td>0</td>
<td>14</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>L2</td>
<td>1.44</td>
<td>3.13</td>
<td>0</td>
<td>15</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>L3</td>
<td>1.38</td>
<td>3.19</td>
<td>0</td>
<td>15</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>L4</td>
<td>1.52</td>
<td>2.96</td>
<td>1</td>
<td>15</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Total (L1–L4)</td>
<td>1.41</td>
<td>3.05</td>
<td>0</td>
<td>15</td>
<td>24</td>
</tr>
<tr>
<td>Proposed (with CA)</td>
<td>L1</td>
<td>0.98</td>
<td>2.26</td>
<td>0</td>
<td>13</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>L2</td>
<td>1.02</td>
<td>2.26</td>
<td>0</td>
<td>13</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>L3</td>
<td>1.08</td>
<td>2.38</td>
<td>0.5</td>
<td>14</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>L4</td>
<td>1.18</td>
<td>2.37</td>
<td>1</td>
<td>13</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Total (L1–L4)</td>
<td>1.07</td>
<td>2.31</td>
<td>1</td>
<td>14</td>
<td>16</td>
</tr>
</tbody>
</table>

All results are depicted as the number of axial computed tomography slices. The number of the last column denotes the number of outliers with an error greater than or equal to 10 slices.

SD, standard deviation; CA, channel attention.
**Supplemental Table S2.** Results of Cross-Validation on 269 Training Data

<table>
<thead>
<tr>
<th>Set no.</th>
<th>Baseline Dice per case</th>
<th>Baseline Dice per slice</th>
<th>Proposed Dice per case</th>
<th>Proposed Dice per slice</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.9595</td>
<td>0.9597</td>
<td>0.9616</td>
<td>0.9619</td>
</tr>
<tr>
<td>2</td>
<td>0.9581</td>
<td>0.9584</td>
<td>0.9593</td>
<td>0.9596</td>
</tr>
<tr>
<td>3</td>
<td>0.9640</td>
<td>0.9640</td>
<td>0.9649</td>
<td>0.9649</td>
</tr>
<tr>
<td>4</td>
<td>0.9636</td>
<td>0.9633</td>
<td>0.9650</td>
<td>0.9648</td>
</tr>
<tr>
<td>5</td>
<td>0.9655</td>
<td>0.9660</td>
<td>0.9663</td>
<td>0.9667</td>
</tr>
<tr>
<td>6</td>
<td>0.9599</td>
<td>0.9601</td>
<td>0.9607</td>
<td>0.9608</td>
</tr>
<tr>
<td>7</td>
<td>0.9629</td>
<td>0.9629</td>
<td>0.9632</td>
<td>0.9632</td>
</tr>
<tr>
<td>8</td>
<td>0.9617</td>
<td>0.9616</td>
<td>0.9633</td>
<td>0.9630</td>
</tr>
<tr>
<td>9</td>
<td>0.9643</td>
<td>0.9652</td>
<td>0.9651</td>
<td>0.9659</td>
</tr>
<tr>
<td>10</td>
<td>0.9638</td>
<td>0.9641</td>
<td>0.9655</td>
<td>0.9654</td>
</tr>
<tr>
<td>Average</td>
<td>0.9623</td>
<td>0.9625</td>
<td>0.9635</td>
<td>0.9636</td>
</tr>
</tbody>
</table>
Supplemental Fig. S1. (A) The network architecture for the localization subtask models. (B) Channel attention block architecture. The number below each block denotes the number of channels. BN, batch norm; ReLU, rectified linear unit; Conv, convolution.
Supplemental Fig. S2. (A) Proposed semi-supervised learning model for the lumbar segmentation method. (B) The architecture of the segmentation network. $L_{CE}$, cross-entropy loss; $Y$, predicted segmentation map; $L_{MScnn}$, Mumford-Shah loss; Conv, convolution; BN, batch norm; ReLU, rectified linear unit.
**Supplemental Fig. S3.** Results of lumbar vertebra (L1, L2, L3, and L4) segmentation from the proposed method on the computed tomography scans of (A) a man and (B) a woman. Dice scores are notated in each result.