Learning Quantitative Calibration of Computed Tomography for Bone Mineral Density Estimation Using Multi-Center Data

Job H. J. Schöller*[1,6], Yi Gu*[1](✉), Keisuke Uemura[2], Masaki Takao[3], Seiji Okada[4], Nobuhiko Sugano[2], Yoshinobu Sato[5], and Yoshito Otake[1](✉)

[1] Division of Information Science, Graduate School of Science and Technology, Nara Institute of Science and Technology, Japan
gu.yi.gu4@naist.ac.jp, otake@is.naist.jp
[2] Department of Orthopeadic Medical Engineering, Osaka University Graduate School of Medicine, Japan
[3] Department of Bone and Joint Surgery, Ehime University Graduate School of Medicine, Japan
[4] Department of Orthopaedics, Osaka University Graduate School of Medicine, Japan
[5] Nara Medical University, Japan
[6] Department of Biomedical Engineering, Eindhoven University of Technology, The Netherlands

Abstract. Opportunistic osteoporosis screening using routine CT scans enables bone mineral density (BMD) assessment without additional imaging modalities. However, conversion from Hounsfield Units (HU) to physical density requires calibration, and scanner-dependent variations limit generalizability across clinical centers. We propose a supervised 3D image-to-image translation framework that learns to synthesize voxel-wise Quantitative Computed Tomography (QCT) density maps of the proximal femur directly from routine CT. For clinical evaluation, volumetric BMD (vBMD) is computed as the mean calibrated density within the proximal femur mask. Agreement is assessed between predicted vBMD and phantom-calibrated reference vBMD using mean absolute error (MAE), Pearson correlation coefficient (PCC), and intra-class correlation coefficient (ICC). We evaluate the method on a large multi-center dataset of 2157 patients from 9 centers, using leave-one-center-out cross-validation. Centers with a similar calibration factors distribution are considered in-distribution (ID), while centers with distinct calibration factors are considered out-of-distribution (OOD), representing a more challenging scenario for model generalization. Performance is compared against conventional calibration baselines. Simply applying the mean calibration factors of the training dataset achieve high agreement on ID data (ICC = 0.981, MAE = 10.0 ± 8.6), but their reliability deteriorates on OOD data (ICC = 0.620, MAE = 40.8 ± 7.4). In contrast, our image-to-image translation model maintains robust agreement on OOD data (ICC = 0.877, MAE = 19.1 ± 10.5), demonstrating resilience to center-specific calibration shifts and improved generalizability across clinical settings.

* Equal contribution.

Keywords: bone mineral density · quantitative CT · image-to-image translation · multi-center generalizability

1 Introduction

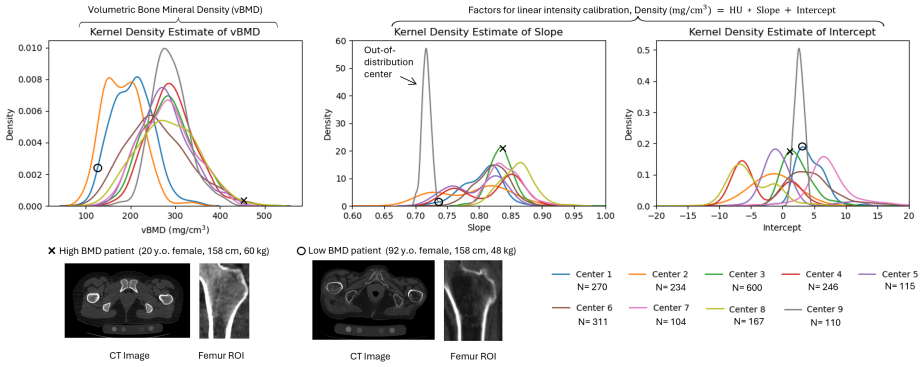

Fig. 1. Distributions of the volumetric bone mineral density (vBMD) and calibration factors (slope and intercept) for the multi-center dataset. Center 9 shows smaller variation in all three parameters and an systematic shift in calibration slope compared to the other centers. and is therefore defined as the "out-of-distribution" center. Two example subjects are highlighted from the dataset, patient X (high BMD female) and patient O (low BMD female).

Osteoporosis is the primary cause of reduced bone density, which increases the risk of fragility fractures [1]. When occurring in the spine or hips, these fractures often lead to physical deformity, functional impairment, and increased mortality [1,2]. Because fracture incidence increases strongly with age, demographic shifts toward older populations are expected to substantially increase the absolute number of osteoporotic fractures worldwide [3]. Early identification of individuals with low bone mineral density is therefore a major clinical objective, as pharmacological therapies have been shown to significantly reduce fracture risk [4,5]. However, many patients are only diagnosed after sustaining a fragility fracture, underscoring the need for scalable and accessible screening strategies [6].

Dual-energy X-ray Absorptiometry (DXA) is currently the gold standard for measuring bone mineral density (BMD) [7]. In addition, Quantitative Computed Tomography (QCT) enables both volumetric and areal measurement of bone mineral density, with volumetric BMD regarded as more sensitive to changes caused by treatment than the two-dimensional area-based measurements provided by DXA [7]. Despite this advantage, conventional QCT is limited by the requirement for an external calibration phantom to convert Hounsfield Units (HU) into physical density (mg/cm^3). This phantom is placed under the patient

during the CT scan and serves as a reference to establish a linear relationship between measured HU values and known physical densities. However, the need for a dedicated phantom restricts QCT to specifically planned examinations, limiting its integration into routine clinical imaging workflows. Opportunistic screening using routine CT scans provides a practical alternative, enabling osteoporosis assessment without the need for a dedicated calibration phantom or additional imaging. Since a large number of patients already undergo CT examinations for various clinical indications, bone mineral density can be evaluated retrospectively from existing scans, without extra radiation exposure, additional costs, or changes to the clinical workflow [8]. This approach enables identification of at-risk individuals who might not otherwise be screened, potentially detecting more patients with low bone density at an earlier stage. Previous studies have investigated deep-learning approaches for BMD regression and automated calibration in this setting [9,10]. However, the methods are validated using single-center data, which do not address the significant variations of calibration parameters in clinical settings. Addressing this variability is essential for developing a clinically robust and widely deployable solution.

In this paper, we propose a method to generate QCT for BMD estimation, specifically targeting the proximal femur region of interest (ROI). We use a large multi-center dataset of 2157 patients from 9 different centers to address the challenges of multi-center generalizability. As illustrated in figure 1, the distribution of volumetric bone mineral density (vBMD) and calibration factors (slope and intercept) exhibits both overlaps and differences. Center 9 in particular, represents a challenge characterized by an offset in calibration slope and a more narrow parameter distribution compared to other centers. Therefore, we make a distinction in the dataset between centers 1 through 8, which represent "in-distribution" (ID) data, and center 9, which represents "out-of-distribution" (OOD) data. By learning to generate voxel-wise QCT maps from routine CT, our approach leverages the strengths of image-to-image translation, a technique widely studied in medical imaging for synthesizing one modality from another while preserving anatomical structure and quantitative information [11]. Unlike other CT-based BMD approaches [9,10,12], our method generates voxel-wise maps directly, preserving spatial density distributions and enabling robust handling of center-specific calibration differences while maintaining accurate bone mineral density assessment. This spatially dense supervision provides vastly richer learning signals than direct regression approaches, which rely on single scalar values and may learn global statistical shortcuts rather than a scanner-normalized representation of physical bone density.

2 Materials and Methods

2.1 Dataset and preprocessing pipeline

For all subjects in the dataset, a CT image was acquired with a calibration phantom positioned beneath the patient. Scanners from GE, Toshiba, Siemens, and Canon are used across centers 1–8, whereas center 9 exclusively uses GE

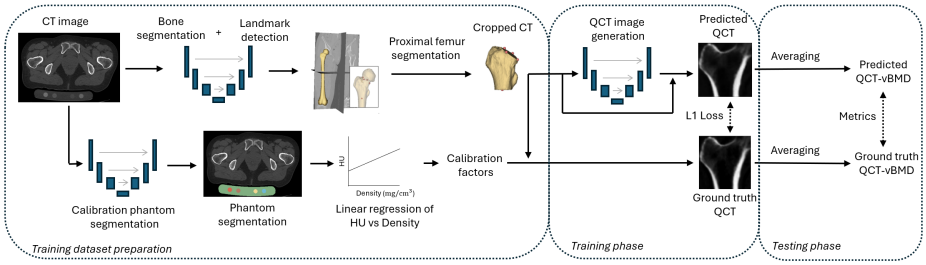

Fig. 2. Schematic overview of the proposed framework. The full pipeline is divided into three stages: training data preparation, model training, and testing. The training data preparation begins with bone segmentation [13] and automated landmark detection [14] to extract the proximal femur ROI from routine CT scans. Calibration factors (slope and intercept) are derived by automatically segmenting the calibration phantom [15] and performing linear regression against known physical densities. These calibration factors are applied to the cropped CT to generate the ground truth (GT) QCT. The image-to-image translation model is trained using L1 reconstruction loss in a residual learning framework, predicting voxel-wise residuals that are added to the input CT to generate the final QCT maps. For evaluation, vBMD is computed as the mean voxel intensity within the proximal femur mask for both GT and predicted QCT volumes.

scanners. Center 9 used a lower-dose acquisition protocol, which may contribute to the observed OOD calibration shift.

In order to train the image generation model on the proximal femur ROI, preprocessing of the full routine CT is necessary (Figure 2). First, a bone segmentation model was applied to isolate the femur [13], followed by a landmark detection model to further localize the proximal femur ROI. The landmark detection model predicts the landmarks by regressing landmark heatmaps using a Bayesian U-Net [14], which is trained on a in-house dataset. Next, to obtain a voxel-wise ground truth (GT), the reference calibration factors are automatically derived for each scan. A dedicated calibration phantom segmentation model is employed [15], followed by linear regression between the known phantom densities and the observed HU values to estimate the calibration slope and intercept for each scan. The GT QCT images are generated by applying the following linear calibration formula to all pixels within the cropped CT volume: $\text{Density}(\text{mg/cm}^3) = \text{HU} \cdot \text{Slope} + \text{Intercept}$. The GT vBMD measurements are computed as the average voxel density within the proximal femur mask. The combination of CT femur bone segmentation, phantom segmentation, and BMD calculation follow the pipeline as described in Uemura et al. [16]. Segmentation errors propagate into the method and are accepted as ground truth. Finally, negative densities in the GT QCT are clipped to 0, and both the GT QCT and cropped CT are resized into fixed volumes of 64x64x64 voxels. The field of view was tightly cropped around the proximal femur, aiming to minimize background tissue while preserving a limited amount of surrounding tissue to support model prediction.

2.2 Image generation model

We formulate QCT generation as a supervised image-to-image translation task, where $X \in R^{H\times W\times D}$ is the input cropped CT volume of the proximal femur, and $Y \in R^{H\times W\times D}$ is the corresponding ground-truth QCT volume obtained via phantom-based linear calibration. The goal is to learn a mapping $f_\theta : X \to \hat{Y}$, parameterized by θ, that minimizes the voxel-wise reconstruction error:

$$\mathcal{L}_{\text{recon}} = \|f_\theta(X) - Y\|_1 \tag{1}$$

In our approach, f_θ is implemented as a simple 5 layer 3D U-Net [17], which is intentionally chosen as a proof of concept to evaluate the proposed formulation. To improve learning stability and preserve voxel-wise HU-to-density relationships, the network is trained in a residual formulation, inspired by residual learning in image denoising [18]. Specifically, the network predicts a residual volume $R_\theta(X)$, defined as the voxel-wise difference between the ground-truth QCT Y and the input CT X, which is then added to X to produce the final QCT estimate:

$$\hat{Y} = X + R_\theta(X) \tag{2}$$

where $\hat{Y}$ denotes the predicted QCT volume. Volumetric bone mineral density (vBMD) is computed as the mean voxel intensity within the proximal femur mask. This scalar value is used for evaluation only and is not part of the training objective (Figure 2).
The implementation is available at: https://github.com/Jobs27/QCT-translation

3 Experiments and results

To ensure a robust evaluation across the multi-center dataset, we employed a leave-one-center-out cross-validation strategy. In this approach, each center is iteratively held out as a test set, while the remaining eight centers are used for model training. After completing all folds, we aggregated the results into two primary categories for analysis.

- In-distribution (ID) Performance: The combined results of test centers 1-8 (N=2047), representing centers with relatively overlapping parameter distributions
- Out-of-distribution (OOD) Performance: The results for test center 9 (N=110), which was identified as a challenge due to its distinct calibration slope and narrower parameter distribution as shown in Figure 1.

ID and OOD categories were determined based on calibration-factor distributions, with ID performance aggregating test results across eight cross-validation folds and OOD performance corresponding to the single center 9 cross-validation fold.

Ethical approval was obtained from the Institutional Review Boards (IRBs) of the institutions participating in this study (IRB approval numbers: 21115 for

Osaka University Hospital, 2021-M-11 for Nara Institute of Science and Technology.) We evaluate the predicted vBMD using the ground truth vBMD, using mean absolute error (MAE), Pearson correlation coefficient (PCC) and intraclass correlation coefficient (ICC) as evaluation metrics. The performance of our image-to-image translation model is compared against four baseline models. The first baseline model, referred to as "No calibration", applies a slope of 1 and an intercept of 0 to the original cropped CT, effectively comparing the raw HU values directly against the ground-truth vBMD. The second baseline, "Mean calibration", computes the average slope and intercept from the training set and applies this calibration to all test subjects. In addition, two regression-based baseline models were implemented using a ResNet18 architecture [19]. Both models take the cropped CT volume as input but differ in their target outputs. The first, "Predicted vBMD", directly regresses a single scalar representing vBMD. The second, "Predicted calibration", predicts two scalars, slope and intercept, which are then applied to the HU values of the cropped CT to compute vBMD. Table 1 provides an overview of all evaluated methods.

Table 1. Overview of the four baseline models and the proposed method, comparing architecture, calibration strategy, and output type. "Computed" QCT applies linear calibration factors to the input CT; "Derived" vBMD is the mean QCT voxel intensity within the proximal femur mask; "Predicted" indicates direct model output.

Method	**Architecture**	**Calibration factors**	**QCT**	**vBMD**
No calibration	–	Slope = 1, intercept = 0	Computed	Derived
Mean calibration	–	Mean slope and intercept	Computed	Derived
Pred. calibration	ResNet18	Pred. slope and intercept	Computed	Derived
Pred. vBMD	ResNet18	–	–	Predicted
Pred. QCT	3D U-Net	–	Predicted	Derived

Table 2 summarizes the performance of the proposed image-to-image translation method compared to the baseline approaches across both ID and OOD datasets. On the ID data, the baseline models achieve a lower error compared to the proposed method, with Predicted calibration having the smallest MAE. The Mean calibration approach shows the highest ICC, while the No calibration baseline shows the highest PCC. The high PCC observed for No calibration reflects the strong linear relationship between Hounsfield Units (HU) and bone mineral density. Since vBMD is obtained through a linear transformation of HU values, applying no calibration preserves a linear correlation, even though the absolute scale is incorrect. PCC only measures the strength of a linear relationship, without accounting for absolute agreement, whereas ICC quantifies both consistency and agreement. Therefore, ICC is a more relevant metric for evaluating vBMD predictions in this setting.

When evaluated on the OOD center, performance of the baseline models deteriorates, where the MAE increases and the ICC drops sharply. In contrast, our proposed QCT prediction model demonstrates better generalizability. Although

Table 2. Model-wise performance for in-distribution (ID) and out-of-distribution (OOD) data, with MAE reported in mg/cm^3. Highlighted are the best metrics for each category. Mean calibration appears to be a sufficient approach for achieving high reliability in ID data (ICC = 0.981, MAE = 10.015 ± 8.609), while No calibration shows the highest PCC (0.985), reflecting the preserved linear relationship between HU and vBMD despite incorrect scaling. Pred. QCT model is the only approach to maintain clinical reliability predicting the OOD center (ICC = 0.877, MAE = 19.145 ± 10.504), whereas No calibration again shows the highest PCC (0.998), driven by linear correlation rather than agreement.

	In-distribution (N=2047)			Out-of-distribution (N=110)		
Model	MAE ± std	PCC	ICC	MAE ± std	PCC	ICC
No calibration	53.820 ± 16.923	**0.985**	0.749	110.504 ± 17.090	**0.998**	-0.157
Mean calibration	10.015 ± 8.609	0.984	**0.981**	40.752 ± 7.378	**0.998**	0.620
Pred. calibration	**9.746 ± 10.334**	0.979	0.978	47.047 ± 9.754	0.995	0.536
Pred. vBMD	12.765 ± 11.075	0.971	0.969	44.898 ± 13.891	0.971	0.550
Pred. QCT	15.244 ± 12.065	0.974	0.957	**19.145 ± 10.504**	0.979	**0.877**

it shows a slightly higher error on ID data, it is the only approach that remains stable when faced with the OOD challenge of Center 9. It achieves a low MAE of 19.1 ± 10.5 (which corresponds to a mean absolute percentage error (MAPE) of 6.49 ± 3.29) and maintains a high ICC of 0.877, whereas the second-best performing baseline on OOD data only achieves an ICC of 0.620 and a MAE of 40.8 ± 7.4 (MAPE = 14.08 ± 1.13). As shown in figure 3, the Pred. QCT predictions align more closely with the identity line for the OOD center compared to the baseline models, which suffer from a consistent underestimation of density. These results suggest that the proposed method is more resilient to center-specific calibration shifts than the baseline approaches.

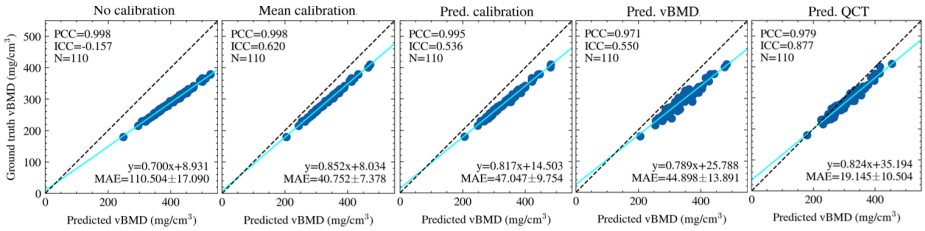

Fig. 3. Comparison of vBMD prediction accuracy across the different models for out-of-distribution (OOD) data (Center 9, N=110). Scatter plots illustrate the relationship between ground truth vBMD and predicted vBMD. The No calibration baseline shows a high PCC due to the linear relationship between HU and bone density, and the Mean, Pred. calibration, and Pred. vBMD baselines exhibit systematic underestimation and low reliability (ICC $\leq$ 0.620) when evaluated on the OOD data. In contrast, the proposed Pred. QCT model (image-to-image translation) demonstrates robustness, maintaining a higher consistency with the identity line and achieving the highest reliability (ICC = 0.877) and lowest error (MAE = 19.145 ± 10.504).

Ablation study. To evaluate the contribution of each design choice, we conducted a systematic ablation study by modifying learning rate, loss functions, and target formulation (direct QCT prediction versus residual learning). Performance metrics were calculated across the entire dataset (N=2157) using the leave-one-center-out cross-validation strategy. The results are summarized in Table 3. The Baseline configuration was defined using a low learning rate ($1e-5$), L1 loss, and direct QCT prediction. Our experimental protocol involved changing one parameter at a time; if an adjustment resulted in improved performance, that variable was maintained for subsequent experiments while a new parameter was modified. As shown in Table 3, increasing the learning rate (Exp. A) improved all 3 metrics. Using this improved learning rate, we found that switching to an L2 loss (Exp. B) or incorporating a Gradient Correlation (GC) loss (Exp. D) did not yield better results. However, shifting the target formulation from direct QCT prediction to residual learning (Exp. C) provided a small improvement, achieving the highest performance with an MAE of 15.4 ± 12.0 , a PCC of 0.968 and an ICC of 0.956. Consequently, the final settings implemented for the proposed framework are those established in Experiment C.

Table 3. Ablation study of the proposed method. Each row modifies one component relative to the previous configuration. The settings resulting in the highest performance are highlighted in Experiment C, which utilizes a $1e-4$ learning rate, L1 loss, and residual target prediction.

Exp.	LR	Loss	Target	MAE ± std	PCC	ICC
Baseline	1e-5	L1	direct QCT	23.047 ± 19.123	0.936	0.882
A (LR)	1e-4	L1	direct QCT	16.411 ±12.129	0.965	0.952
B (Loss)	1e-4	L2	direct QCT	18.064 ± 13.064	0.961	0.942
C (Residual)	1e-4	L1	residual QCT	**15.443 ± 12.019**	**0.968**	**0.956**
D (GC)	1e-4	L1 + GC	residual QCT	15.669 ± 11.797	0.966	0.955

4 Summary

We propose a supervised 3D image-to-image translation framework for phantomless bone mineral density (BMD) estimation from routine CT. The proposed model learns to synthesize voxel-wise Quantitative Computed Tomography (QCT) maps of the proximal femur. Ground-truth QCT volumes are generated using phantom-based linear calibration, and the network is trained with an L1 reconstruction objective.

The method is evaluated on a large multi-center dataset (N = 2157, 9 centers) using a leave-one-center-out strategy and is compared against conventional global calibration baselines, including No calibration, Mean calibration, and regression-based calibration models. One center, characterized by a calibration slope shift and narrower parameter distribution, is treated as out-of-distribution to assess robustness under center-specific calibration shifts.

Our results demonstrate that a simple Mean calibration approach is sufficient for achieving high reliability within in-distribution centers. However, this method fails to generalize when faced with significant calibration shifts in the out-of-distribution center, where the measurement reliability drops considerably. In contrast, the proposed image-to-image framework maintains substantially higher agreement and lower error under these conditions, demonstrating improved generalizability across centers. Nevertheless, the proposed framework still exhibits systematic underestimation of BMD in the out-of-distribution setting, which is acknowledged as a limitation of the current approach.

Acknowledgments. This study was supported by the Japan Agency for Medical Research and Development (AMED, JP25hma322015.) and the Japanese Orthopaedic Association (JOA-Subsidized Science Project Research: 2023-2).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Testa EJ, Callanan TC, Evans AR, Aaron RK (2022). Osteoporosis and Fragility Fractures. R I Med J (2013) 105(8):15-21.
2. Johnell O, Kanis JA (2006) An estimate of the worldwide prevalence and disability associated with osteoporotic fractures. Osteoporos Int 17(12):1726–1733. doi:10.1007/s00198-006-0172-4
3. Sing CW, Lin TC, Bartholomew S, Bell JS, Bennett C, Beyene K et al (2023) Global Epidemiology of Hip Fractures: Secular Trends in Incidence Rate, Post-Fracture Treatment, and All-Cause Mortality. J Bone Miner Res 38(8):1064–1075. doi:10.1002/jbmr.4821
4. Cummings SR, San Martin J, McClung MR, Siris ES, Eastell R, Reid IR et al (2009) Denosumab for prevention of fractures in postmenopausal women with osteoporosis. N Engl J Med 361(8):756–765. doi:10.1056/NEJMoa0809493
5. Compston JE, McClung MR, Leslie WD (2019) Osteoporosis. Lancet 393(10169):364–376. doi:10.1016/S0140-6736(18)32112-3
6. Siris ES, Miller PD, Barrett-Connor E, Faulkner KG, Wehren LE, Abbott TA et al (2001) Identification and fracture outcomes of undiagnosed low bone mineral density in postmenopausal women: results from the National Osteoporosis Risk Assessment. JAMA 286(22):2815–2822. doi:10.1001/jama.286.22.2815
7. Kanis JA. Diagnosis of osteoporosis and assessment of fracture risk. Lancet. 2002 Jun 1;359(9321):1929-1936. doi: 10.1016/S0140-6736(02)08761-5.
8. Jang S, Graffy PM, Ziemlewicz TJ, Lee SJ, Summers RM, Pickhardt PJ (2019) Opportunistic Osteoporosis Screening at Routine Abdominal and Thoracic CT: Normative L1 Trabecular Attenuation Values in More than 20 000 Adults. Radiology 291(2):360–367. doi:10.1148/radiol.2019181648
9. Fang Y, Li W, Chen X, Chen K, Kang H, Yu P et al (2021) Opportunistic osteoporosis screening in multi-detector CT images using deep convolutional neural networks. Eur Radiol 31(4):1831–1842. doi:10.1007/s00330-020-07312-8

10. Sollmann N, Löffler MT, El Husseini M, Sekuboyina A, Dieckmeyer M, Rühling S et al (2022) Automated Opportunistic Osteoporosis Screening in Routine Computed Tomography of the Spine: Comparison With Dedicated Quantitative CT. J Bone Miner Res 37(7):1287–1296. doi:10.1002/jbmr.4575
11. Dayarathna S, Islam KT, Uribe S, Yang G, Hayat M, Chen Z et al (2024) Deep learning based synthesis of MRI, CT and PET: Review and analysis. Med Image Anal 92:103046. doi:10.1016/j.media.2023.103046
12. Bartenschlager S, Cavallaro A, Pogarell T, Chaudry O, Uder M, Khosla S et al (2023) Opportunistic Screening With CT: Comparison of Phantomless BMD Calibration Methods. J Bone Miner Res 38(11):1689–1699. https://doi.org/10.1002/jbmr.4917
13. Soufi M, Otake Y, Iwasa M, Uemura K, Hakotani T, Hashimoto M et al (2025) Validation of musculoskeletal segmentation model with uncertainty estimation for bone and muscle assessment in hip-to-knee clinical CT images. Sci Rep 15(1):125. doi:10.1038/s41598-024-83793-7
14. Hiasa Y, Otake Y, Takao M, Ogawa T, Sugano N, Sato Y (2020) Automated Muscle Segmentation from Clinical CT Using Bayesian U-Net for Personalized Musculoskeletal Modeling. IEEE Trans Med Imaging 39(4):1030–1040. doi:10.1109/TMI.2019.2940555
15. Uemura K, Otake Y, Takao M, Soufi M, Kawasaki A, Sugano N et al (2021) Automated segmentation of an intensity calibration phantom in clinical CT images using a convolutional neural network. Int J Comput Assist Radiol Surg 16(11):1855–1864. doi:10.1007/s11548-021-02345-w
16. Uemura K, Otake Y, Takao M, Makino H, Soufi M, Iwasa M et al (2022) Development of an open-source measurement system to assess the areal bone mineral density of the proximal femur from clinical CT images. Arch Osteoporos 17(1):17. doi:10.1007/s11657-022-01063-3
17. Ronneberger O, Fischer P, Brox T (2015) U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF (eds) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Lecture Notes in Computer Science, vol 9351. Springer, Cham, pp 234–241. doi:10.1007/978-3-319-24574-4_28
18. Zhang K, Zuo W, Chen Y, Meng D, Zhang L (2017) Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising. IEEE Trans Image Process 26(7):3142–3155. doi:10.1109/TIP.2017.2662206
19. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, pp 770–778. doi:10.1109/CVPR.2016.90