**Peer Review File**

Article information: https://dx.doi.org/10.21037/jmai-23-10

**Reviewer comments**

**Comment 1:**

Summary:

The paper presents an application of deep learning for approximating femoral neck BMD from native X-Ray images of hand, knee, and pelvis. The fusion model solution combines image and co-variate data to produce results that are then compared to three different baseline models. The authors provide further insight to some of the hyperparameter choices and to the contribution of different co-variate data variables in the supplementary. Testing the method with images from different bone sites is interesting and the manuscript is concisely written. However, the novelty of the study is somewhat limited, as the problem has been studied in several papers before, especially using chest or pelvis x-rays. The used dataset is unfortunately small in a deep learning context, and therefore raises doubts about the results and their significance. Anyhow, the results themselves are not that impressive, and hardly justify the bold conclusions the authors made about the superiority of the approach, its feasibility for opportunistic screening, and the potential to change the current standard of care for osteoporosis.

**Reply 1:**

The novelty of the current dataset is that it is estimating BMD at the FN using x-rays of body parts (knee, hand) that do not include the femur. We agree, the dataset is small, which is why we used k-fold cross validation. The fact that we can show superior performance (even with a smaller margin) compared to the baselines shows the significance and efficiency of our approach.

**Changes in the text:**

We have changed the claim from “Opportunistic screening for low BMD from conventional x-ray can change the current standard of care for osteoporosis” to “Opportunistic screening for low BMD from conventional x-ray could help address the known care gap in osteoporosis management, i.e., its under-screening and under-diagnosis.”

**Comment 2:**

Major points:

- the level and significance of the results was exaggerated
- reference and comparison to previous limilar studies is lacking
- the variance and inconsistency in the results indicates that the dataset is too small for the deep learning task at hand
- the authors did not use an external test set. In addition, they did not explain if the

test set was used only in the final phase after all hyperparameters and trained weights were fixed or was the method tuned based on the results from the test set. It sounds like the latter is the case since authors, for example, tell that the T-score threshold was chosen based on the ROC curve. This would mean that the model was fitted to the test data, and it remains unclear whether the method will generalize well if the model is used with a new dataset.

**Reply 2:**
Regarding the first 3 points, they are reiterated and our responses are described in the more detailed remarks and questions, below (i.e., Replies 4, 6, 7, 8, below).

Regarding the 4th point, please note that the model was trained using the k-fold cross-validation approach with a held-out test set. This means that all the post-training model tuning work, such as hyperparameter optimization, choosing the thresholds, etc., were done by evaluating the performance of the validations set of each fold. In particular, we chose the hyper parameters in a way to minimize the absolute difference in BMD between the model and ground truth in the validation sets (not the test set). Therefore, the final reported results on the test set have not contributed to any parameter choice. Subsequently, the T-Score thresholds were chosen based on the ROC curves that are computed using the validation split of each fold in order to provide results on sensitivity and specificity at a given threshold, however, the AUC is threshold independent.

**Changes in the text:**
Added the following to the last paragraph on page 7 for better clarification: "The training set goes through a k-fold cross validation process, and the validation splits are used for final model tuning."

**Detailed remarks and questions:**

**Comment 3:**
P.2,l.46-9: "...from 553 unique patients (51% male) in patients with osteoarthritis, aged between 48 to 83 years old." The sentence structure here sounds odd: from patients in patients.

**Reply 3:**
We agree with the reviewer and have corrected the sentence.

**Changes in the text:**
We have changed the text to read "…from 553 unique patients with osteoporosis (51% male), aged between 48 to 83 years old."

**Comment 4:**

P.3 highlight box:
The claimed implication of changing the current standard of care for osteoporosis is an overstatement. it may be difficult to demonstrate clinical value for presented results of predicting bone mineral density, which in turn only moderately predicts fragility fractures. I would suggest a little more conservative claims of implications. Developing this kind of method can enable opportunistic side diagnosis if successful, but changing the standard of care would require something more.

**Reply 4:**
We agree with the reviewer and have modified the implications text.

**Changes in the text:**
We have changed the claim from "Opportunistic screening for low BMD from conventional x-ray can change the current standard of care for osteoporosis" to "Opportunistic screening for low BMD from conventional x-ray could help address the known care gap in osteoporosis management, i.e., its under-screening and under-diagnosis."

**Comment 5:**
P.4,l.80:
The recommendations for screening are not entirely clear-cut. For example, WHO states in their report (although not very recently) that "widespread screening at the menopause on the basis of BMD alone is not generally recommended because of the poor sensitivity and specificity of BMD measurement". Many studies have questioned the fracture prediction ability of both BMD and FRAX. There is definitely value in the opportunistic screening of low bone density, but this claim of screening recommendations is a simplification of a more complicated matter.

**Reply 5:**
We appreciate the reviewer's insights that the value of BMD and FRAX is a complicated matter. We believe that the screening recommendations are clear for the US based on the reference cited (USPTF). We have expanded the description of the recommendation to which we referred. We have also added text to clarify that the value of an opportunistic screen of low BMD from x-ray is not necessarily the BMD value *per se*, but rather an alert to a referring clinician of their patient potentially at risk for fracture, thus encouraging the clinician to conduct a fracture risk assessment.

**Changes in the text:**
We added a sentence before "Despite the recommendations (7)" to say "The United States Preventative Screening Task Force (USPTF) recommends BMD testing by DXA in women 65 and older, and younger women with certain clinical risk factors". We also added a final sentence to the paragraph: "An opportunistic

screen for low BMD from x-ray could alert a care provider to conduct a clinical fracture risk assessment, and refer for DXA if appropriate."

**Comment 6:**
P.4,l.87-90:
"...but a simple model without need for manual feature extraction, that could be implemented opportunistically has yet to be developed." The authors should review and explain previous literature more thoroughly. For example, a quick search gives several studies with similar research question and study setup than the current study.
For example, see:
Hsieh CI et al. 2021: Automated bone mineral density prediction and fracture risk assessment using plain radiographs via deep learning
Sato et al. 2022: Deep Learning for Bone Mineral Density and T-Score Prediction from Chest X-rays: A Multicenter Study
Zhang et al. 2020: Deep learning of lumbar spine X-ray for osteopenia and osteoporosis screening: A multicenter retrospective cohort study
Chen et al. 2021: Application of deep learning neural network in predicting bone mineral density from plain X-ray radiography

**Reply 6:**
We concur with the reviewer. The original text was highlighting the relative paucity of studies that have predicted femoral neck BMD from x-rays that do not include the femur, i.e., the hand x-ray studies described. We have added the recent publications that you have noted, highlighting that other research often aims to predict BMD of the body part imaged, though one group predicted femoral neck and lumbar BMD from chest x-rays, and another group extended the clinical utility by offering fracture risk prediction.

**Changes in the text:**
We have added the following text: "The majority of these have used pelvic (22, 23) or lumbar (24) x-rays, and have shown utility in predicting BMD of those body parts, given that osteoporosis is monitored at those sites. Other researchers have had success predicting BMD from chest x-rays.(25) A large study recently extended the utility of a deep-learning algorithm to include fracture risk assessment (23). The current study shows similar results when estimating femoral neck BMD from the pelvic x-rays, but also estimates femoral neck BMD and T-Score from x-rays of body parts that do not include the femur, namely the hand and knee. Fewer studies have estimated BMD from hand x-rays. "

**Comment 7:**
P.9, l.200:
"Notably, all the knee models show better performance than the other body parts".

On what results is this conclusion based? Table 4. shows DL AUC 0.92/0.88 for Pelvis where as table 3. shows DL AUC 0.87/0.89 for the knee. I interpret that result as being better for the pelvis than for the knee.

**Reply 7:**
This was by analytically and visually interpreting the plots shown in Figure 5. All the knee models can estimate the ground truth T-score (agree with the ground truth T-scores) better than the other body parts, judging by the slope and the intercept of the fitted linear model, as well as the computed 95% confidence intervals. However, the reviewer is right about the pelvis model showing a "classification" performance when the AUC metric is considered. While the interpretation about the knee model can be attributed to having more data points, the better classification of the pelvis model can be due to the fact that the pelvis X-rays contain images of the femoral neck, which will have a substantial contribution to the BMD estimate of the femoral neck.

**Changes in the text:**
We have added to section 3.2, after line 226: "The pelvis model had the best classification performance, which may be due to the fact that the pelvis x-rays the femoral neck, which will have a substantial contribution to the BMD estimate of the femoral neck, whereas the knee and hand x-rays do not."

**Comment 8:**
P.9, l.205:
"Furthermore, while the agreement mean (blue line in the figure) is almost the same for all the baselines and the DL model, it is clear that the baseline model predictions are more biased than the DL model."
I do not see how it is clear that the baseline predictions are more biased. There seems to be slightly bigger variance, but the mean difference in BMD is close to zero and I do not see significantly stronger proportional bias in the baseline either.

**Reply 8:**
To our understanding, proportional bias is observed when the difference in values resulting from two methods increases/decreases in proportion to the mean values. It is visually observed in the baseline plots (bottom 3 rows) of Figure 6 that the plotted data points are scattered over a downward line with a slope higher (more biased) than that of the DL models (top row). This can also be interpreted by the CIs being narrower in the DL (top row) models plots. However, the reviewer could be right about the strong tone of the claim, as there is "slight" bias in the baseline models.

**Changes in the text:**
The sentence has been reworded to: "it is clear that the baseline model predictions

are slightly more biased”

**Comment 9:**
P.10, l.214:
The aim is to predict the classification label of low BMD (T-Score < -1), i.e. osteopenia, but here it states that "The optimal T-Score cut-off (lower false positive rate vs. higher true positive rate) are calculated directly from the ROC curve of each model". This may artificially produce better results because the threshold that happens to work best for this data is selected. I believe the threshold of low bone density should be fixed to T-Score < -1 without any further calibration.

**Reply 9:**
The T-Score thresholds were chosen based on the ROC curves that are computed using the validation split of each fold in order to provide results on sensitivity and specificity at a given threshold, however, the AUC is threshold independent. The T-Score from the validation data was applied to the held-out test set, thus is not the threshold that works best for the test set.

**Changes in the text:**
We have modified the last 2 paragraphs of section 2.2 as follows: “T-Scores were derived from BMD values using female peak bone mass from NHANES III (14). Area under the receiver operating curve (ROC) was employed to assess the “low BMD” (yes/no DXA T-score <-1) classification performance. Algorithm-derived (predicted) T-Score thresholds were chosen based on the ROC curves computed using the validation split of each fold.
The predicted BMD of the test set was calibrated using the k-fold cross-validation technique (15), and the T-Scores were derived as described above. In the test sets, model performance at estimating continuous BMD vs. DXA ground truth were assessed with Passing-Bablok (16) and Bland-Altman (17). Area under the receiver operating curve (ROC), accuracy, sensitivity, and specificity were employed to assess the “low BMD” (DXA T-score <-1) classification performance at the algorithm-derived T-Score thresholds obtained from the validation splits. Given the sex difference in the prevalence of low BMD (2), performance was assessed in males and females, separately. The R package cutpointr (18) was used to calculate the 95% confidence intervals on the AUC using 4,000 bootstraps using in-bag values in the AUC_b column of the bootstrap results.”

**Comment 10:**
P.11, l.232:
"it is demonstrated that imaging data contains rich diagnostic information which is not readily discernible to the human eye but can be extracted with modern machine learning approaches.". Below a certain level, low bone density can be

detected by an experienced radiologist. I wasn't shown in this study that a human doctor could not perform a similar approximation, and there was no reference to any other studies that can back up this claim.

**Reply 10:**
We appreciate the reviewer's concern and agree that we should have a reference to support this claim. Osteopenia is not detected on conventional radiographs until 20 to 40 percent of bone mass has been lost. This is now referenced in the text. At 30%, this corresponds to a T-Score of -2.1 for the FN, and a T-Score of-3.0 for L1-4. As such, an algorithm that can identify the earlier stage of demineralization, at a T-score of -1, offers an advantage to the human eye.

**Changes in the text:**
We have revised the sentence and added the reference: "This study showcases the potential of modern machine learning in identifying patients with low BMD from routinely acquired x-rays, even when the femoral neck is not included in the field of view. By comparing to baseline models, this study demonstrates that imaging data contains rich diagnostic information that can be extracted with modern machine learning approaches. A radiologist can detect osteopenia on conventional radiographs only when 20-40% of bone mass has been lost (19). As an example, if someone had a BMD of 0.858 g/cm2 at the femoral neck (which is the mean BMD for a female aged 20-29 years from NHANES III (14)), then a loss of 30% of bone mass would correspond to a T-Score of -2.1 (assuming the use of a female reference population as recommended by WHO (20)). Similarly, if someone had a BMD of 1.064 g/cm2 at L1-L4 (which is the mean BMD for a female aged 20-29 years from NHANES (21)), then a loss of 30% would correspond to a T-Score of -3.0. As such, an algorithm that can identify the earlier stage of demineralization, at a T-score of -1, offers an advantage to the human eye."

**Comment 11:**
P.11, l.235:
"Strengths of the current method and study include the performance of the algorithm without the need for segmentation or manual feature extraction, and the availability of DXA BMD as the ground truth."
This is of course relative and a matter of opinion, but I found the mediocre performance of the algorithm a weakness of this study. When we are not predicting the actual outcomes of the disease but merely a proxy of the bone density which is only one predictor of fractures, the expectation for accuracy is higher to be of any clinical relevance. Another thing with this sentence: isn't "the availability of DXA BMD as the ground truth" a requirement for the supervised machine learning approach rather than a strength of the study?

**Reply 11:**

We agree with the reviewer and have removed the availability of DXA as ground truth (though we note that other research on estimating low BMD from hand x-rays has not used DXA as ground truth, as we discuss in the manuscript). We have instead described another strength of the study as the algorithm's ability to estimate BMD at the femoral neck, even when the femoral neck is not in the x-ray analyzed (i.e., in x-rays of the knee and hand). We understand the reviewer's opinion that the accuracy should be higher to be of clinical relevance, but we respectfully argue that would be true if this was a test ordered by a clinician in order to assess BMD. Rather, the idea of this algorithm is that anyone getting an x-ray could have their x-ray analyzed at the time of acquisition, and those suspected of having low BMD could have a clinical fracture risk assessment. If such an assessment led to DXA referral, the DXA would provide the accuracy (gold standard) that the reviewer is suggesting. The difference is the idea of opportunistic screening (taking advantage of the wealth of x-ray data) in order to prioritize patients who would benefit from additional assessment. We have added a sentence to section 4.4 to clarify this point.

**Changes in the text:**
**Section 4.2:** "Strengths of the current method and study include the performance of the algorithm without the need for segmentation or manual feature extraction, and its ability to estimate BMD at the femoral neck, even when the femoral neck is not in the x-ray analyzed (i.e., in x-rays of the knee and hand). This study had several limitations..."
**Section 4.5:** "While the DL model will be unlikely to rival the accuracy and precision of DXA in the quantification of low BMD, an opportunistic screening approach is particularly attractive because it can help identify and prioritize patients who are currently overlooked. Osteoporosis has a known care gap.(5) It is prevalent, silent, and preventable with treatment."

FIGURES:
**Comment 12:**
Figure 1:
I suggest putting "patients" and "x-ray/DXA pairs" as titles for the figures so that it becomes obvious even without reading the caption. Also, in the text of the manuscript it is mentioned that T=0 window was used for test sets but in figure 1 (and also figure 2) I get the impression that T=24 refers to both training and test set.

**Reply 12:**
The validation and the test splits have t = 0. The title of Figures 1 and 2 have been edited to show this, and to clarify patients and x-ray/DXA pairs.

**Changes in the text:**

The titles and axes labels in Figures 1 and 2 have been modified.

**Comment 13:**
Figure 5
The scales for the X and Y axis are different, which distorts the scatter plot. I do not see any reason for this as the BMD scale should be the same be it predicted or ground truth.

**Reply 13:**
We had originally thought these scales enabled better visualization of the scatter plot and we had provided the identity line (y=x) to facilitate comparison between the x- and y-axes. In response to the reviewer's request, however, we have modified the axes scales to mach.

**Changes in the text:**
See Figure 5.

**Comment 14:**
Figure 6:
The X-axis title "Mean BMD" might confuse some readers not familiar with Bland–Altman analysis. It could be clarified where this mean comes from.

**Reply 14:** We have clarified both the y-axis and x-axis titles, to be "Difference in BMD (Predicted BMD - the Ground Truth DXA BMD)", and "Mean BMD (Derived from the Model and from the Ground Truth (DXA))", respectively.

**Changes in the text:**
See Figure 6.

TABLES
**Comment 15:**
Table 1:
The order of different sites in the table is different from, for example, figures 5 and 6. I suggest using the same order of bone sites consistently throughout the manuscript.

**Reply 15:**
We agree and have reordered the figures, tables and text to all have the same order (hand, knee, pelvis).

**Changes in the text:**
Any mentions of these body parts in the text have been re-ordered, as described above.

**Comment 16:**
Tables 2-4:
The purpose of the bold font here is unclear. I assume it is to emphasize the best result in every row, but at least the bolding of CIs and some of the false negatives and false positives seems inconsistent.

A few things pop up from the results that would be good to ponder in the discussion. Many models are producing very different results for males and females. For example, DL AUC for pelvis in females is 0.88 compared to 0.78 in males where as the knee model works better for males and the hand model is again better for females. Same inconsistency can be observed in the baseline models. One could argue that all this is random variation caused by the too-small dataset. In the test set for the pelvis, there are <50 samples, which can cause very unstable results. This is also observed in the wide confidence intervals.

How were the confidence intervals calculated in the results? Although maybe obvious for the author, there are many ways to do it and it would improve transparency if it was explained in the methods or in the supplementary.

**Reply 16:**
We have removed bolding from CIs and FN/FPs. DL AUC for pelvis in females is 0.88 compared to 0.92 in males. The reviewer likely meant the Accuracy was lower in males (0.78) vs females (0.88). The CIs of the accuracies overlap, which makes it difficult to draw conclusions from this. We believe that the AUC estimates are less impacted by the sample sizes than the accuracy, sensitivity and specificity obtained after applying a threshold. As such, we believe it is better to compare performance by sex using AUC. Notably, the AUC is similar between sexes in all 3 body parts. We have added the method of calculating CIs to the methods section.

**Changes in the text:**
We added: "The R package cutpointr (18) was used to calculate the 95% confidence intervals on the AUC using 4,000 bootstraps using in-bag values in the AUC_b column of the bootstrap results."

SUPPLEMENTARY
**Comment 17:**
The ablation study is a good addition to the paper. But, if it was to justify the choices for the most important hyper-parameters, it could mention the reason why InceptionV3 architecture was chosen. What about did it work with random initialization of the weights or only with pre-trained weights? Also, some more details on how the model training was performed would provide useful insight. For example, were all the layers of the pretrained model trained or just some of them?, How did the learning curve converge?

The supplementary could include more information on how the statistical analysis of the results was done (my previous point about confidence intervals).

**Reply 17:**
InceptionV3 was chosen among a few commonly used image backbones in the computer vision literature, specifically those with proven success in the medical domain (see reference #6 in our supplementary material document). Furthermore, we leveraged the transfer learning technique to train our model, mainly due to the limited size of our dataset that would prevent us from training a model from scratch. It is also shown in multiple computer vision related research papers and experiments that the neural network models with weights initialized from the pretrained weights on the ImageNet dataset train more smoothly than those with random weights initialization. For more details about our training strategy, please refer to the newly added section titled "Choice of the Image Backbone Architecture and the Training Strategy Details" in the supplementary material document.

**Changes in the text:**
A new section titled "Choice of the Image Backbone Architecture and the Training Strategy Details" was added to the supplementary material document to explain the architecture selection procedure and our training strategy.