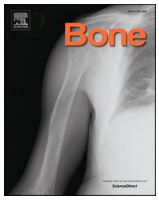
Bone 179 (2024) 116987

Contents lists available at ScienceDirect

Bone

journal homepage: www.elsevier.com/locate/bone



Full Length Article

The effect of deep learning-based lesion segmentation on failure load calculations of metastatic femurs using finite element analysis

Ali Ataei [a,*], Florieke Eggermont [a], Nico Verdonschot [a,c], Nikolas Lessmann [b], Esther Tanck [a]

[a] Orthopaedic Research Lab, Radboud university medical center, P.O. Box 9101, 6500, HB, Nijmegen, the Netherlands
[b] Diagnostic Image Analysis Group, Department of Medical Imaging, Radboud university medical center, Nijmegen, the Netherlands
[c] Laboratory for Biomechanical Engineering, University of Twente, Enschede, the Netherlands

ARTICLE INFO

Keywords:
Bone fracture risk
Femur
Finite element
Deep learning
Lesion segmentation

ABSTRACT

Bone ranks as the third most frequent tissue affected by cancer metastases, following the lung and liver. Bone metastases are often painful and may result in pathological fracture, which is a major cause of morbidity and mortality in cancer patients. To quantify fracture risk, finite element (FE) analysis has shown to be a promising tool, but metastatic lesions are typically not specifically segmented and therefore their mechanical properties may not be represented adequately. Deep learning methods potentially provide the opportunity to automatically segment these lesions and change the mechanical properties more adequately. In this study, our primary focus was to gain insight into the performance of an automatic segmentation algorithm for femoral metastatic lesions using deep learning methods and the subsequent effects on FE outcomes. The aims were to determine the similarity between manual segmentation and automatic segmentation; the differences in predicted failure load between FE models with automatically segmented osteolytic and mixed lesions and the models with CT-based lesion values (the gold standard); and the effect on the BOne Strength (BOS) score (failure load adjusted for body weight) and subsequent fracture risk assessments.

From two patient cohorts, a total number of 50 femurs with osteolytic and mixed metastatic lesions were included in this study. The femurs were segmented from CT images and transferred into FE meshes. The material behavior was implemented as non-linear isotropic. These FE models were considered as gold standard (Finite Element no Segmented Lesion: FE-no-SL), whereby the local calcium equivalent density of both femur and metastatic lesion was extracted from CT-values. Lesions in the femur were manually segmented by two biomechanical experts after which final lesion segmentation for each femur was obtained based on consensus of opinions between two observers. Subsequently, a self-configuring variant of the popular deep learning model U-Net known as nnU-Net was used to automatically segment metastatic lesions within the femur. For these models with segmented lesions (Finite Element with Segmented Lesion: FE-with-SL), the calcium equivalent density within the metastatic lesions was set to zero after being segmented by the neural network, simulating absence of load-bearing capacity of these lesions. The models (either with or without automatically segmented lesions) were loaded incrementally in axial direction until failure was simulated. Dice coefficient was used to evaluate the similarity of the manual and automatic segmentation. Mean calcium equivalent density values within the automatically segmented lesions were calculated. Failure loads and patterns were determined. Furthermore, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were calculated for both groups by comparing the predictions to the occurrence or absence of actual fracture within the patient cohorts.

The automatic segmentation algorithm performed in a none-robust manner. Dice coefficients describing the similarity between consented manual and automatic segmentations were relatively low (mean 0.45 ± standard deviation 0.33, median 0.54). Failure load difference between the FE-no-SL and FE-with-SL groups varied from 0 % to 48 % (mean 6.6 %). Correlation analysis of failure loads between the two groups showed a strong relationship ($R^2 > 0.9$). From the 50 cases, four cases showed clear deviations for which models with automatic

Abbreviations: FE, Finite Element; FE-no-SL, Finite Element no Segmented Lesion; FE-with-SL, Finite Element with Segmented Lesion; PPV, Positive Predictive Value; NPV, Negative Predictive Value; BOS, BOne Strength; QCT, Quantitative CT (computed tomography); FOV, Field of View; HU, Hounsfield Unit; BW, Body Weight.

* Corresponding author.
E-mail address: Ali.Ataei@radboudumc.nl (A. Ataei).

https://doi.org/10.1016/j.bone.2023.116987
Received 21 July 2023; Received in revised form 29 November 2023; Accepted 4 December 2023
Available online 5 December 2023

lesion segmentation (FE-with-SL) showed considerably lower failure loads. In the whole database including osteolytic and mixed lesions, sensitivity and NPV remained the same, but specificity and PPV decreased from 94 % to 83 %, and from 78 % to 54 % respectively from FE-no-SL to FE-with-SL.

This study indicates that the nnU-Net yielded none-robust outcomes in femoral lesion segmentation and that other segmentation algorithms should be considered. However, the difference in failure pattern and failure load between FE models with automatically segmented osteolytic and mixed lesions were relatively small in most cases with a few exceptions. On the other hand, the accuracy of fracture risk assessment using the BOS score was lower compared to the FE-no-SL. In conclusion, this study showed that automatic lesion segmentation is a none-solved issue and therefore, quantifying lesion characteristics and the subsequent effect on the fracture risk using deep learning will remain challenging.

1. Introduction

Bone ranks as the third most frequent tissue affected by cancer metastases, following the lung and liver. The majority of bone metastases are caused by breast and prostate cancer [1]. The incidence rate of bone metastases in cancer patients ranges from 35 % to 73 % [2]. Bone metastases are classified as osteolytic, osteoblastic, or mixed, based on how they primarily interfere with normal bone remodeling. Osteolytic metastases are usually characterized by destruction of normal bone while osteoblastic metastases are characterized by deposition of new bone [1]. Bone metastases are often painful, and may result in impaired mobility and pathological fracture [1,3,4]. Metastases are most frequently seen in the axial skeleton. In the appendicular skeleton, metastatic lesions in the proximal femur are the most common [5], which is probably associated with the well-developed vascular system within the femur allowing cancer cells to enter the femur from the original tumor site.

Pathological fractures are a major cause of morbidity and mortality in cancer patients with metastases [1,3]. Approximately 60 % of the pathological fractures occur in the femur [6]. Due to its specific mechanical functionality, the proximal femur needs accurate consideration to estimate the fracture risk. Patients may receive chemotherapy, conservative radiotherapy, undergo preventive surgery, or a combined treatment in order to reduce fracture risk, mitigate pain and/or reduce lesion growth [7,8].

To quantify fracture risk, finite element (FE) analysis has shown to be a promising tool [7–9]. FE models take location and geometrical characteristics of the lesion and bone into account, as well as the general bone quality, whereas clinical radiographical measurements on 2-D or 3-D images do not typically consider these features [8]. In two patient studies, we showed that an FE-based fracture risk assessment was more accurate than the current Dutch clinical guidelines (30 mm cortical involvement) (sensitivity: 100 % versus 86 %, specificity: 74 % versus 42 %, positive predictive value (PPV): 39 % versus 19 %, and negative predictive value (NPV): 100 % versus 95 %) [7,8]. Recently, we started to clinically implement this method under the name BOne Strength (BOS) score. This is an easy-to-use score based on FE outcomes [9] which assists physicians to determine the best treatment together with the patient.

Despite the advantage of using FE models over clinical measurements that are recommended by the current guidelines, there is still room for improvement in the models. For example, the mechanical bone properties (stiffness and strength) in FE models are assumed to be solely dependent on calcium equivalent density (which is related to bone mineral density (BMD))of the local bone. However, it is questionable whether empirical relationships between CT density and mechanical properties obtained with healthy bone maintain their validity for densities within metastatic lesions. Osteoblastic lesions typically show a higher density, but are suspected to be weaker than the CT-value would imply [10], although Kaneko indicated that under compressive loads CT-based predictions could still maintain their validity for osteoblastic bone [11]. Due to this uncertainty, application of CT-based bone strength calculations using the BOS score has been limited to femurs affected with predominantly osteolytic lesions. Furthermore, osteolytic lesions show low CT density values, but their actual strength may even be lower (or negligible) than the density would suggest [10]. Generating a reliable FE model in which the mechanical properties of the metastatic lesions are appropriately represented requires a reliable segmentation of these lesions, allowing for alternative mechanical constitutive laws between density and mechanical properties.

It is possible to segment the metastatic lesions using manual segmentation techniques on radiographical images. However, this is a highly tedious task and the obtained segmentations are dependent on the training and experience of the person who is doing the segmentations. Previously, we studied the reliability of manual segmentations of femoral metastatic lesions and confirmed that manual segmentation resulted in variable segmentation results [3]. This indicates that manual segmentation of bone metastases is not reliable and should probably not be used for FE modeling.

Utilizing current deep learning algorithms, automatic segmentation can now potentially be enabled, saving time and likely providing a more robust way to segment deviations of density distributions (lesions) from normal [12]. Neural networks are widely used to recognize abnormalities through learning patterns related to unhealthy tissue. They have shown to be promising to detect, for example, mammographic lesions [13], spinal metastatic lesions [14–16], or they are used in digital pathology [17]. A neural network known as nnU-Net showed strong performance across 53 diverse tasks including organ, tissue, and tumor segmentation [18], and is potentially capable to detect bone metastatic lesions. Application to metastatic lesion detection is potentially hampered by the fact that manual reproducibility annotation is low [3], but it is suggested that despite these noisy segmentations in the training data, deep learning methods can result in robust segmentations [19].

In this study, we aimed to use a deep neural network to segment femoral osteolytic lesions and the osteolytic part of mixed lesions, followed by adjusting the calcium equivalent density within the segmented lesions to zero, hypothesizing that this method would deliver similar (or slightly lower) bone strength results compared to the FE models without lesion segmentation (the gold standard) as osteolytic lesions hardly have bone density. Our primary focus was to gain insight into the challenges associated with the automatic segmentation of femoral metastatic lesions using deep learning methods and the subsequent effects on FE outcomes. Hence, in this study we evaluated our new approach, and formulated the following research questions: 1) what is the similarity between consented manual segmentation and automatic segmentation, 2) what is the difference in failure pattern and failure load between FE models with automatically segmented osteolytic and mixed lesions, and the models with CT-based lesion values (the gold standard), and 3) how do the possible differences between the two methods translate into a different BOne Strength (BOS) score and subsequent fracture risk assessments (validating the BOS score for the extreme-case scenario in which calcium equivalent density within the metastatic lesions was set to zero).

2. Methods

2.1. Patients and CT scans

CT scans of 57 femurs with femoral osteolytic and mixed metastatic

lesions were used to train the neural network (see Section 2.3). Another group of 50 femurs (35 osteolytic and 15 mixed femurs from two patient cohorts [7,8]) was included for the FE simulations which was also used for comparison of the FE model with the current guidelines [8]. All CT scans were obtained from four institutes within the Netherlands (ethical approval was obtained from all participating centers). Patients had various cancer types such as multiple myeloma, breast, lung, prostate, thyroid, sarcoma, non-Hodgkin's lymphoma, and kidney, and received radiotherapy between 2006 and 2009 (cohort I [7]), and between 2015 and 2017 (cohort II [8]). Patients were followed up for a period of six months after CT scanning to monitor whether they sustained a fracture or died within that period of time.

The scanning protocol enabling calibration to calcium equivalent density has previously been published [20]. Briefly: Every patient underwent a quantitative CT scan (QCT). A solid calibration phantom containing four known calcium equivalent densities (50, 100, 150, 200 $\frac{mg}{cm^3}$) was scanned along with the patient at the level of the proximal femur. The QCT had the following settings: 120 kVp, 220 mA, slice thickness of 3 mm, pitch 1.5, spiral and standard reconstruction, field of view (FOV) 480 mm, and in-plane resolution 0.9375 mm.

2.2. Deep learning framework

One of the neural network architectures specifically developed for biomedical image segmentation is the U-Net [21], a convolutional neural network with a U-shape. The "U" shape of the network architecture indicates the symmetrical structure which enables the network to capture local and global features of an input image at multiple scales while retaining the spatial information. Consequently, U-Net can adapt to various segmentation tasks, being capable of operating on input images with various size and resolution, and combining features from different spatial regions of the input image. In our study, we used a variant of this network known as nnU-Net [18], which is capable of configuring itself and adapts to the dataset automatically. This method was used to implement automatic dataset preprocessing, configuration of network architecture parameters and training hyperparameters, and post-processing for our dataset of femurs containing osteolytic metastatic lesions. The training process utilized the SGD optimizer, with an initial learning rate of 1e–2, a weight decay of 3e–5, a momentum of 0.99, and spanned across 1000 training epochs. The loss function was the sum of cross-entropy and Dice loss. Each CT scan in the training dataset ($n = 57$) was divided into left and right femur sub-volumes. The image volumes of femurs with osteolytic metastases were used as input to the neural network, together with their corresponding manual segmentation (true labels, see Section 2.3 for details on the manual segmentation process). The output of the nnU-Net automatic configuration and training process were fully trained models, which were applied to the femurs of our patient cohort ($N = 50$) to generate automatic osteolytic lesion segmentations. Network training was based on five-fold cross-validation [22] so that the automatic segmentation for each femur was obtained from a network which was not trained with that femur. Any network output outside of the femoral bone boundary was automatically removed using the femoral segmentation mask (see Section 2.4). No further modifications were done on the outcome segmentation of the neural network.

2.3. Training lesion dataset

Femoral osteolytic lesions were selected out of 57 cancerous femurs (43 % in the femoral head, 17 % in the diaphysis, and 40 % in both areas). Lesions in the femur were manually segmented using Mimics (version 14.0, Materialise, Leuven, Belgium) by two biomechanical experts trained by a musculoskeletal radiologist. The segmentations were variable between the two experts (Dice score around 0.5 [3]). Therefore, final lesion segmentation for each femur was obtained based on consensus of opinions between two observers, followed by segmentation modification if required. During the segmentation, the following four rules were used by the experts [3]: First, lesions with altered cancellous bone were included in the segmentations. Second, missing bone, especially in diaphyseal osteolytic lesions, was segmented. The normal thickness of the bone relative to slices proximal and distal to the segmented slice and to the contralateral femur was compared. Third, in case of mixed lesions, only the osteolytic part of the lesions was considered, segmented and included in the training dataset. Finally, the sclerotic rim of osteolytic lesions was excluded from the segmentation.

2.4. FE models

Patient-specific 3-D models of the 50 femurs were obtained by segmenting each femur in CT images using an automatic method (https://grand-challenge.org/algorithms/femur-segmentation-in-ct/). Briefly, this method uses two convolutional neural networks, of which one segments the bone and the other labels them into left and right femur. Subsequently, the resulting 3-D surface mesh was converted to a solid mesh of tetrahedral elements. The Hounsfield units (HU) of the CT scan were used to determine the patient-specific material properties for each element [20]. For this, a mean diaphyseal slice calibration, using the calibration phantom, was performed to convert the HUs to calcium equivalent densities. The calcium equivalent densities were subsequently used to calculate the ash densities and the isotropic non-linear material properties of the bone [23].

The material behavior was implemented as non-linear isotropic where the post-failure material behavior for each element was represented by an initial perfectly plasticity phase, followed by a strain softening phase, and finally an indefinite perfectly plastic phase [23]. For each FE simulation, a cup was placed on the head of the femur and the model was distally fixed at the knee joint center by two bundles of high-stiffness (200 $\frac{MN}{m}$) springs [8]. This alignment was chosen to mimic the stance position. During the FE simulation (MSC.MARC 2021.4, MSC Software Corporation, Santa Ana, CA, USA), the model was loaded in axial direction by increments of 0.1 mm. It was assumed that failure was reached when the force-displacement curve had a clear peak or when the simulation did not numerically converge any further. The BOS score was calculated by dividing the strength of the femur (i.e. the failure force) by the body weight (BW) of the patient (in N).

For the FE models that were considered as gold standard, the calcium equivalent density of both femur and metastatic lesion was extracted from CT-values [8]. For the approach with segmented lesions, the calcium equivalent density within the metastatic lesions was set to zero after being segmented by the neural network. For ease of reading, the first group models were named as "Finite Element no Segmented Lesion" (FE-no-SL), and the second group as "Finite Element with Segmented Lesion" (FE-with-SL).

2.5. Outcome measure and analysis

We calculated Dice coefficients to evaluate the similarity of the manual and automatic segmentation. A Bland-Altman plot and correlation analysis were used to demonstrate the difference and correlation between the failure loads of the FE-no-SL and FE-with-SL groups. The impact of the lesion size on potential differences between the two methods was investigated and visualized as correlation with the failure loads differences. Mean calcium equivalent density values within the automatically segmented lesions were calculated. The plasticity status, i.e. which phase of the non-linear material behavior the elements were (elastic phase, initial perfectly plastic phase, strain softening phase, indefinite perfectly plastic phase), obtained from FE simulation of the femurs in FE-no-SL group was visually compared to those of FE-with-SL group to investigate the difference in failure patterns with respect to the lesion location. Failure loads corrected for the patients' body weight

were calculated and reported under the name BOS score. Finally, the clinical translation of the obtained results was assessed by calculating the sensitivity, specificity, PPV, and NPV of the BOS score for which a threshold equal to the highest sum of sensitivity and specificity of the FE-no-SL group was considered. These clinically relevant values were compared between the FE-no-SL and FE-with-SL groups. Results were separated and visualized in two groups: the whole dataset comprising osteolytic and mixed lesions ($n = 50$), and the dataset with exclusively the osteolytic lesions ($n = 35$).

3. Results

The automatic segmentation algorithm performed in a none-robust manner. Dice coefficients quantifying the resemblance of manual and automatic segmentations were relatively low (mean 0.45 ± standard deviation 0.33, median 0.54). It was noticed that there was a number of cases with relatively small lesions, which were not adequately identified and segmented. This resulted in very low Dice coefficients, which in turn lowered the average Dice coefficient. Additionally, slice-based comparison between manual segmentation and neural network segmentation showed that osteolytic lesions located at the femoral head or around the trochanter level, were segmented more comparable to the manual segmentation (example in Fig. 1: upper row). Contrarily, lesions located distally, and along the femoral bone diaphysis, were sometimes segmented with more deviations from the manual segmentation, indicating that the segmentation algorithm performed worse in the diaphyseal area as compared to the metaphyseal region (Fig. 1: lower row).

3.1. Failure load correlation between FE-no-SL and FE-with-SL groups

FE-no-SL group failure loads ranged from 2839 N to 10,965 N (mean 6295 N), whereas failure loads obtained from FE-with-SL group ranged from 2124 N to 10,965 N (mean 6047 N). Hence, the mean of the failure load of the FE-with-SL was lower, which is to be expected as in those cases the calcium equivalent density of the lesion is set to zero, whereas in the FE-no-SL the calcium equivalent density remained unchanged relative to the CT-values. Mean calcium equivalent density values within the automatically segmented lesions ranged from 0 to 880 mg/cm^3. Failure load difference (with or without automatic segmented lesions) varied from 0 % to 48 % (mean 6.6 %). Correlation analysis of failure loads between the FE-no-SL and FE-with-SL groups showed a strong relationship ($R^2_{\text{osteolytic \& mixed}} = 0.93$, $R^2_{\text{osteolytic}} = 0.9$, Fig. 2).

Four patients, containing only osteolytic lesions, indicated in Fig. 2 (P1-4), had the highest difference in failure load between FE-no-SL and FE-with-SL models ($\text{diff}_{P1} = 48$ %, $\text{diff}_{P2} = 40$ %, $\text{diff}_{P3} = 30$ %, and $\text{diff}_{P4} = 26$ %). This is also visible in the Bland-Altman plot (Fig. 3) where these patients clearly deviate from the average failure load difference line. These four patients can be considered as outliers as they were placed beyond the upper limit line (mean difference + 1.96 standard deviation). In all four patients, the models with automatic lesion

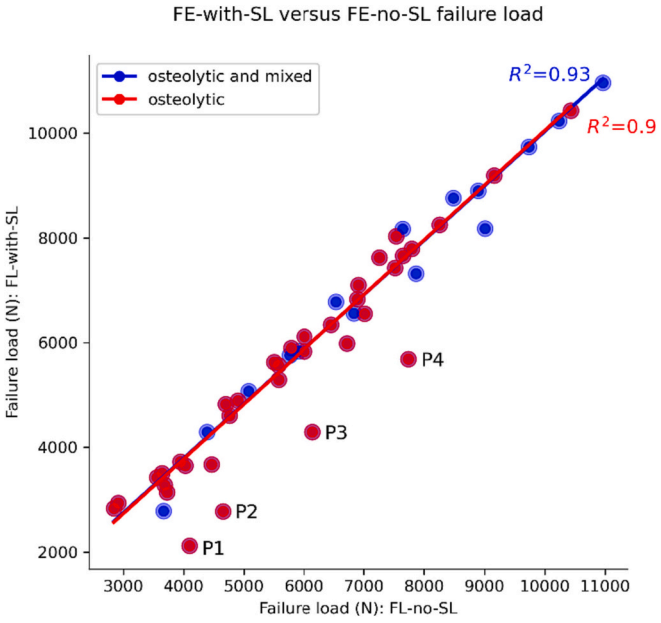


Fig. 2. The correlation between failure loads obtained from femurs in the FE-with-SL group, and femurs in the FE-no-SL group for osteolytic and mixed lesions ($n = 50$; red and blue dots) and osteolytic lesions ($n = 35$; red dots). P1–4 are patients clearly deviating from the regression lines. This figure should be printed in color. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

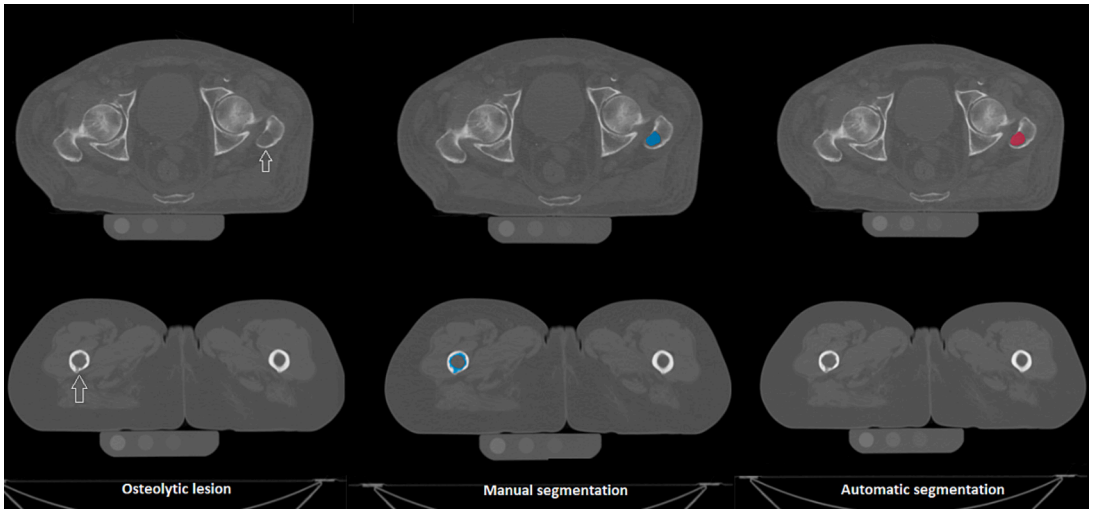


Fig. 1. Osteolytic lesion and segmentation visualization. Upper row shows an osteolytic lesion in the greater trochanter (white arrow) and the manual segmentation (mask in blue) which matches with the automatic segmentation (mask in red). This example only compared the manual and automatic segmentation and showed the ability of the neural network to adequately segment the metastatic lesion within the femoral bone. Lower row shows another osteolytic lesion in the diaphysis (white arrow) and the manual segmentation (mask in blue). This example indicates a deficiency of the neural network to segment the metastases lesion within the femoral bone (no red mask). This figure should be printed in color. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

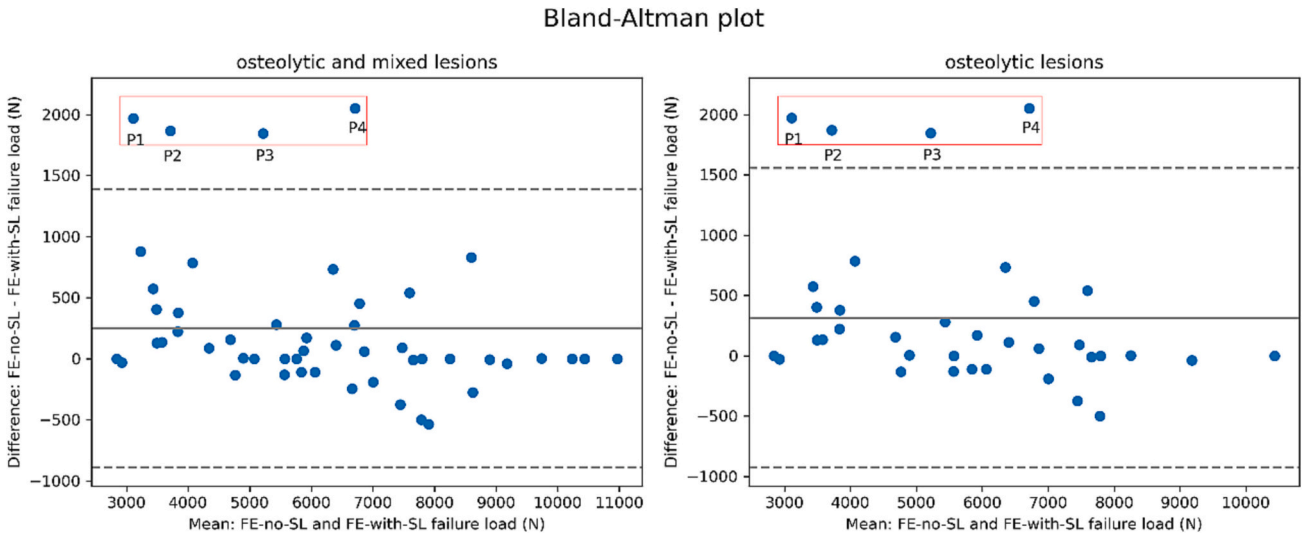


Fig. 3. Bland-Altman plot representing the failure load difference between the FE-no-SL and FE-with-SL groups against the mean failure load for patients with osteolytic and mixed lesions (left), and patients with osteolytic lesions (right). Red boxes include femurs with the most deviating failure load difference from the average line (P1-4). This figure should be printed in color. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

segmentation (FE-with-SL) showed considerably lower failure loads. In three of these four cases, patients had low-contrast osteolytic lesions located around the trochanter and bone diaphysis which were segmented and set to zero-density. In the remaining case, osteolytic lesions around the bone diaphysis were segmented excessively large and therefore a relatively large volume of calcium equivalent density was set to zero.

Lesion size ranged from 0 to 44,321 mm^3 (mean 9904 mm^3). The correlation between failure load difference and lesion size showed that, overall, there was a trend that bigger segmented lesions (of which the whole lesion volume calcium equivalent density was set to zero) led to larger differences of failure load between the two models (Fig. 4). However, there was a large variation around this trend suggesting that not only size of a lesion is determining strength, but that other parameters (such as location of the lesion, actual density on the CT scan) are importantly affecting calculated strength as well.

3.2. Plasticity status difference between FE-no-SL and FE-with-SL groups

In general, plasticity regions within the femurs were at similar locations. As an example, the correspondence in plasticity pattern between FE-no-SL and FE-with-SL models is visualized for two patients (Fig. 5). In both patients, the highest plastic deformation occurred around the lesion area.

3.3. Failure load and clinical translation

BOS scores (i.e. failure loads corrected for patients' body weights) were calculated and shown for the whole dataset comprising osteolytic and mixed lesions, and for only the osteolytic lesions of the dataset (Fig. 6). Sensitivity, specificity, PPV and NPV were calculated for both groups (Table 1). In the whole database including osteolytic and mixed lesions, sensitivity and NPV remained the same, but specificity and PPV decreased by 11 % and 24 % from FE-no-SL to FE-with-SL, respectively. In the osteolytic subgroup, also the sensitivity and NPV of FE-with-SL were similar to the FE-no-SL group, but specificity and PPV were 15 % and 26 % lower compared to FE-no-SL, respectively.

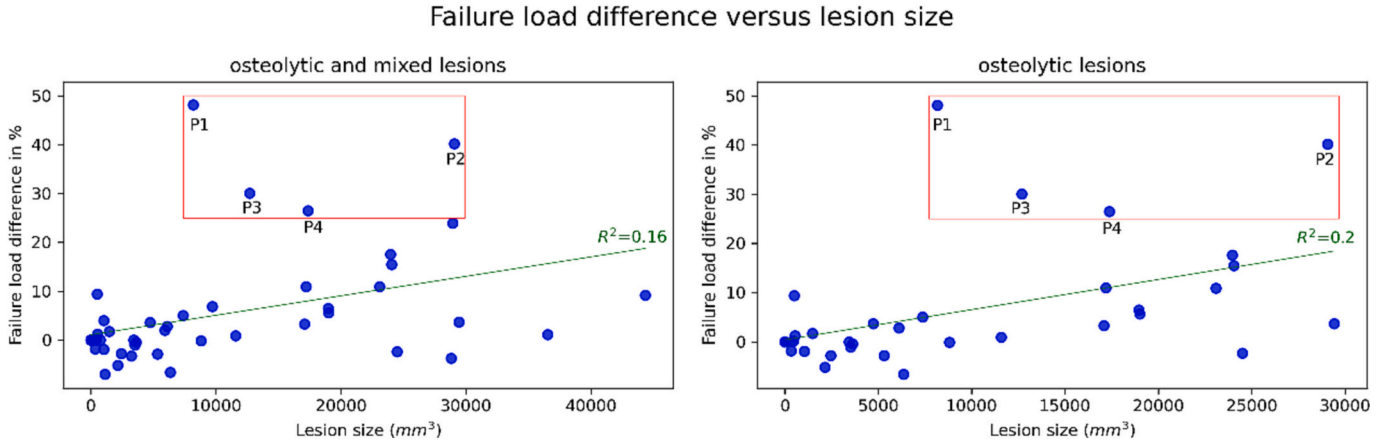


Fig. 4. The correlation between failure load difference (100 % * (FL-no-SL minus FE-with-SL)/FE-no-SL) and lesions size of the osteolytic and mixed group (left), and osteolytic group (right). Red boxes include femurs with the most deviating failure load difference from the regression line (P1-4). The green lines indicate linear regression. This figure should be printed in color. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

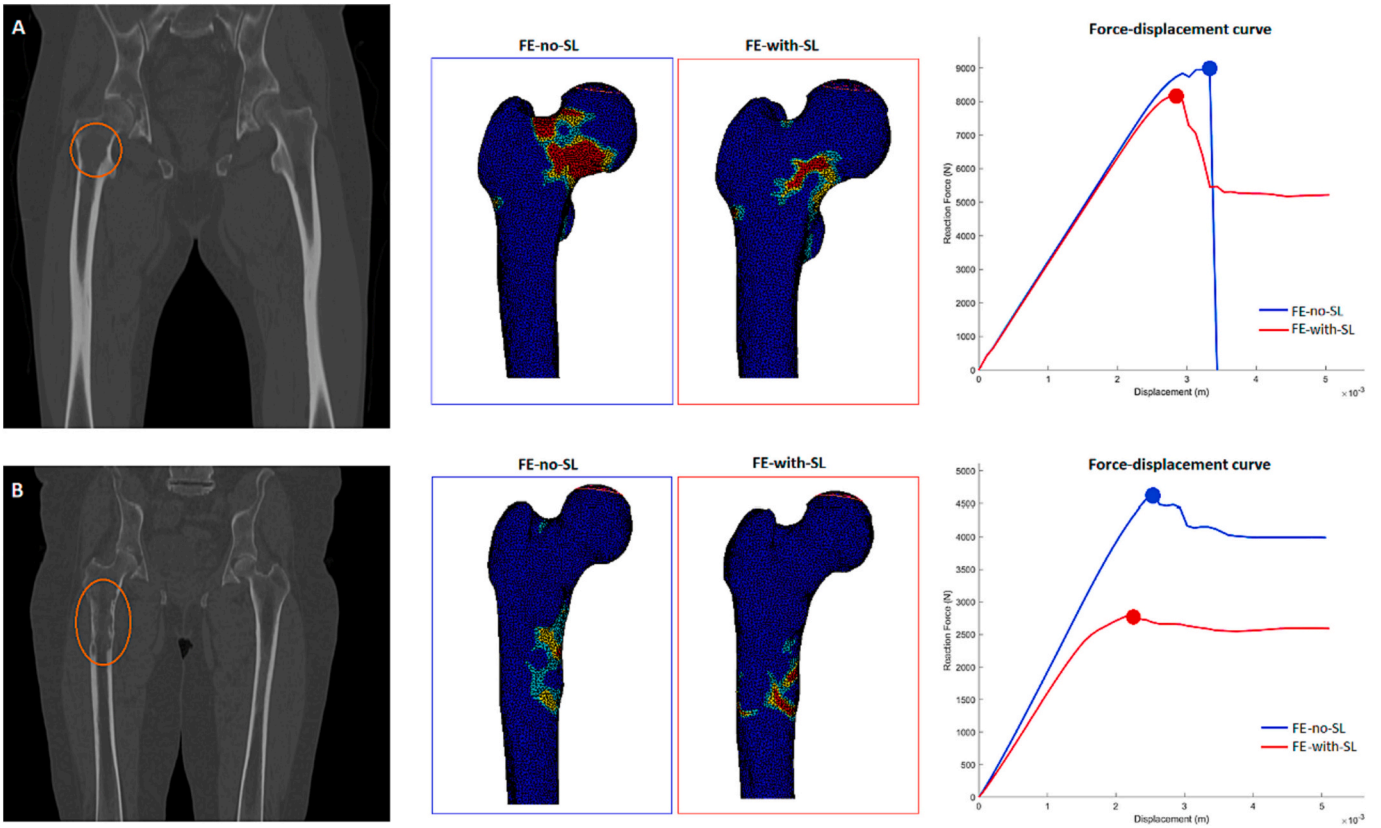


Fig. 5. Femoral osteolytic lesions of two patients are shown (coronal view - orange circle). Failure patterns (plasticity status) obtained from FE-no-SL model (blue box) and FE-with-SL model (red box) are visualized. The plasticity status is shown at the maximal load (red and blue dot) in the Force-displacement curve (right hand side). This figure should be printed in color. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4. Discussion

In this study, our primary focus was to gain insight into the performance of an automatic segmentation algorithm for femoral metastatic lesions using deep learning methods and the subsequent effects on FE outcomes. Therefore, we aimed to determine the similarity between manual segmentation and automatic segmentation and observe how this affected the FE analysis as a tool to calculate bone strength in cancer patients with femoral bone metastases. Others have used similar techniques and proposed combinations with machine learning techniques to enhance clinical applicability of these tools [24]. We utilized a combination of FE with a neural network to automatically segment femoral metastatic lesions, of which we subsequently manipulated the lesion density within the FE model (FE-with-SL), and investigated whether it resulted in similar bone strength compared to that of FE models without lesion density manipulation (FE-no-SL - the gold standard). Finally, the potential impact of the differences was assessed by translating the findings towards the changes in BOS score.

This study shows that automatic lesion segmentation of metastatic lesions in the femur is highly challenging. The expected robustness as suggested by Rolnick et al. [19] when noisy annotations are present in the training set, was not observed in this study. Due to the irregularity of the femoral metastatic lesions particularly in terms of shape, size and density-appearance, generating a high-quality training dataset with an acceptable reliability level is a tedious and time-consuming task [25]. Therefore, probably when using an annotated approach to learn a neural network, the training set requires more and diverse lesion patterns [26–30], particularly where cortex involvement is apparent. For example, lesions located in the bone shaft typically involved cortical bone (both outer and inner surfaces). Segmentation of a lesion intersecting the femoral bone boundary appeared to be more difficult to learn by the neural network than the lesions without any intersection (as for example typically occurring in the trabecular area of the proximal femur). Therefore, the outcome of the neural network on segmentations of proximal femur lesions was more robust compared to femoral shaft lesions. This problem can be partly due to our imbalanced training dataset containing osteolytic lesions in the femoral head or diaphysis region (43 % in the femoral head, 17 % in the diaphysis, and 40 % in both areas). Before this method can be applied clinically, these issues need to be improved and the algorithm needs to better perform in terms of its robustness and validity.

Other deep learning frameworks than the nnU-Net were considered in this study. We preliminary tested other U-net network architectures, but found that the nnU-Net [18] showed the best performance in our preliminary studies. In addition, nnU-Net as a self-configuring network which can automatically implement dataset preprocessing, post-processing, and network parameters adaption is well-ranked in medical image segmentation challenges. Nevertheless, the applicability of the nnU-Net to validly handle diverse and complex datasets as utilized in this study and its performance remains uncertain and it is doubtful whether it will generate robust and reliable lesion segmentation results. Another option to automatically segment the lesions would be to use a labeling method, in which mainly on image-level (a bone contains either metastatic lesions or not) instead of segmentation-level the labels are assigned. This method has also been used in histopathology and it was shown that by using a few segmentation-level labels with a majority of image-level labels the network performed comparable to a fully annotated segmentation dataset [31].

When looking at the effect of the segmentations on the FE outcomes, generally, there was a strong correlation between the FE models without segmentations and those with automatically segmented lesions. However, in four patient cases (P1-P4) large deviations in failure load were

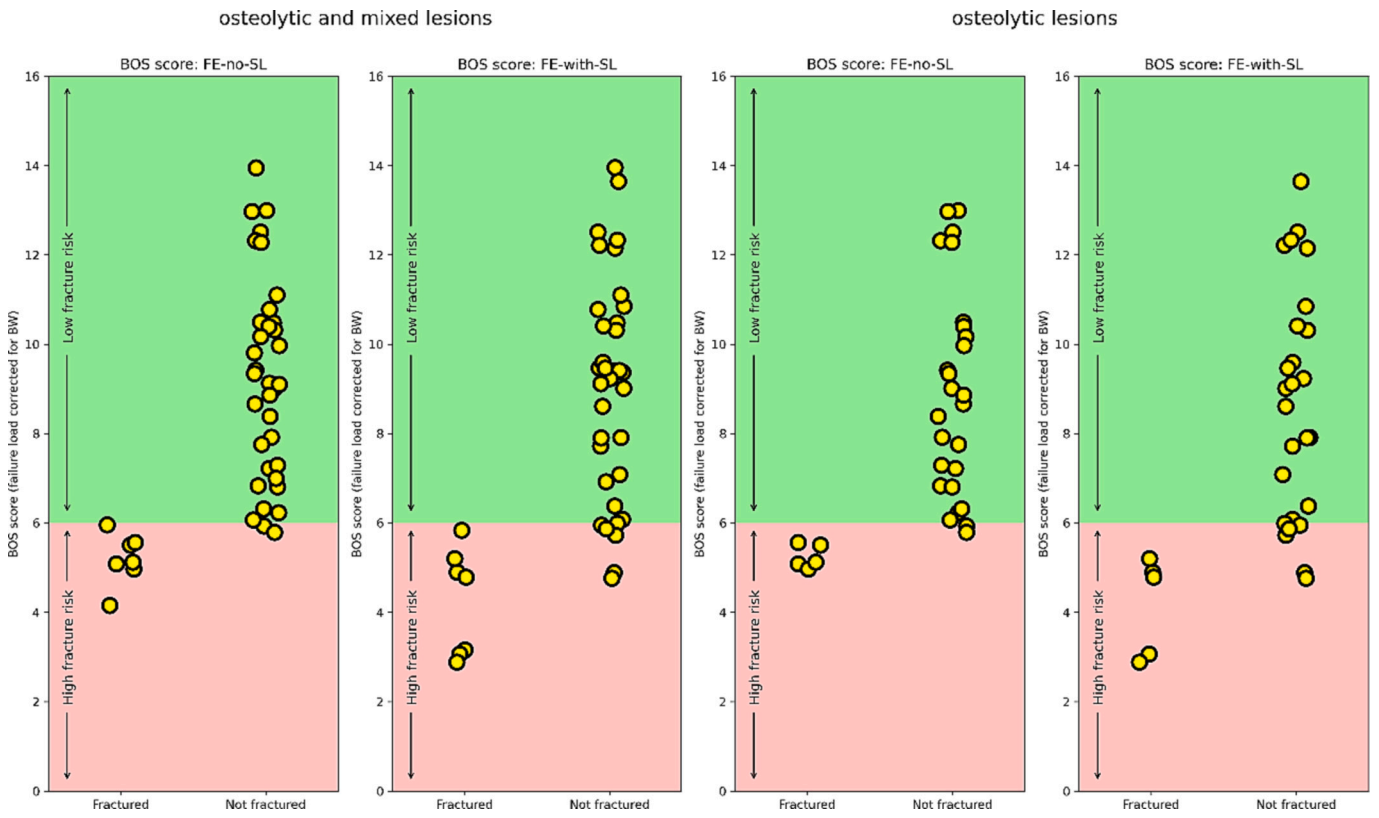


Fig. 6. Each yellow dot represents a femur in the database. The graph provides comparison between FE-no-SL and FE-with-SL BOS scores (failure loads corrected for patients body weight [9]) of the patients with osteolytic and mixed lesions, and osteolytic lesions, respectively. Each graph is divided to a high fracture risk area (in pink) and a low fracture risk area (in green) based on a threshold equal to highest sum of sensitivity and specificity of the FE-no-SL group values. "Fractured" and "Not fractured" columns are based on actual clinical observations. This figure should be printed in color. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) of the FE-no-SL and FE-with-SL models, for the whole dataset (n = 50) and the osteolytic metastases in the dataset (n = 35). The threshold is equal to highest sum of sensitivity and specificity values. The outliers (patients P1-4) were included in the analyses of all the femurs (containing both types of osteolytic and mixed lesions), and the femurs containing only osteolytic lesions. The smaller size of the osteolytic only group may have resulted in a larger effect of the outliers on the diagnostic accuracy values.

	Osteolytic and mixed		Osteolytic	
	BOS FE-no-SL (threshold = 6)	BOS FE-with-SL (threshold = 6)	BOS FE-no-SL (threshold = 6)	BOS FE-with-SL (threshold = 6)
Sensitivity	100 %	100 %	100 %	100 %
Specificity	94 %	83 %	92 %	77 %
PPV	78 %	54 %	71 %	45 %
NPV	100 %	100 %	100 %	100 %

found between the two methods. It obviously is essential to understand the underlying reasons in these four cases before this method can be applied clinically. Further detailed analyses on the four cases revealed that P1 (48.13 % failure load difference) contained low-contrast osteolytic lesions with relatively high CT-values. P2 (40.2 % failure load difference) contained multiple small osteolytic lesions around the proximal femur which were segmented by the network as one large lesion. P3 (30.06 % failure load difference) contained multiple osteolytic lesions around the bone diaphysis which were clearly segmented excessively by the algorithm. P4 (26.51 % failure load difference) contained multiple low-contrast osteolytic lesions around the bone diaphysis for which all the elements inside the segmented part were set to zero strength. Hence, it appears that the automatic segmentation tool had difficulties to segment multiple and low-contrast lesions, which complies with literature [25]. Furthermore, relatively low Dice coefficients were found which was probably caused by uncertainties in the manual segmentations (ground truth) and the subsequent automatic segmentation results. This also explains the large difference in failure load for these four cases as 'low-contrast' means 'relatively high CT-value'. Hence, in a CT-value-based calculation, these lesions get rather high stiffness and strength properties, whereas in the automatically segmented models the calcium equivalent density in the lesions is set to zero with very low stiffness and strength properties.

Correlation of failure load difference with lesion size (Fig. 4) showed that the difference in failure load was generally larger for larger lesion sizes. The failure load difference of the relatively small lesions (lesion size$<$10000mm^3) was smaller than 10 %, showing that a small lesion size may decrease the impact of the lesion location (i.e. a lesion in a critical area). Additionally, we observed that the small size of these lesions minimally influenced the final calculated failure loads in those cases. Therefore, we argue that the reported low mean Dice coefficient has minimal impact on some cases, but future improvements are needed for the segmentation of both osteolytic and osteoblastic lesions. Furthermore, it was found that larger lesions did not necessarily result in weaker bone strength, indicating that other factors such as location and shape of the lesion are of importance for bone strength as well [32–34]. The significance of the lesion size is amplified when clinicians generally tend to rate a bone with larger lesion as a weaker bone [35], but this study shows that this is not always true.

Obviously, this study has further limitations. An important limitation of our study was that we set the calcium equivalent density in the

segmented lesions to zero to simulate bone weakness within the lesions. This may be a logical assumption as osteolytic lesions will exhibit degraded mechanical properties. However, Stadelmann et al. [10] showed little change of tissue properties of osteolytic metastatic bone within the spine. Whether these findings are also valid in the femur is unknown, but it does indicate that reducing the load bearing capacity of osteolytic lesions to zero should be regarded as the extreme effect on femoral bone strength. Another aspect is the fact that bone material properties were assumed to be linear elastic-plastic and had no poro-elastic dependent behavior. Whyne et al. [36], validated a burst fracture model of the metastatic vertebra, showing indeed that a poro-elastic finite element model can realistically represent failure of a metastatic vertebra. More recently, Costa et al. [37], opted for modeling both intact bone and osteolytic lesions as isotropic, elastic-plastic material, although they did not validate the obtained simulation results with experimental measurements. As far as the authors are aware of it remains unclear whether femoral bone strength is sensitive to the use of poro-elastic or linear elastic-plastic modeling of the metastatic lesions; more research is required to investigate this and its clinical relevance. Hence, further investigation on the consequences of metastatic lesions in bone on the remaining mechanical properties and its load-bearing capacity is needed. This issue becomes even more important for osteoblastic lesions. Calcium equivalent density values are typically high in osteoblastic lesions, which may lead to overestimation of the strength in that region affecting the overall strength of the bone. To address the possibility of underestimation or overestimation of the failure load, it is necessary to investigate and assign a more accurate metastatic bone density-elasticity mechanical behavior to the FE models, as suggested by Kaneko et al. [11].

Another limitation is that we do not know what the bone strength of these patients was, but defined a BOS score to grade strength against clinical failure risk. Clinically it is not feasible to obtain strength data other than following the patients and record whether they develop a fracture or not; this is what we did in our clinical studies [7–9]. Furthermore, it should be realized that the CT scans are taken at one time-point and that metastatic lesions may be subject to change (either become smaller due to treatment or progressively grow). This situation is identical for the CT-based models as well as for the models with automatically segmented lesions and does not influence the comparative nature of the study, but may influence the actual failure risk for the patients after the CT scans are made.

5. Conclusion

This study indicates that the nnU-Net yielded none-robust outcomes in femoral lesion segmentation and that other segmentation algorithms should be considered. However, the difference in failure pattern and failure load between FE models with automatically segmented osteolytic and mixed lesions were relatively small in most cases with a few exceptions. On the other hand, the accuracy of fracture risk assessment using the BOS score was lower compared to the FE-no-SL. In conclusion, this study showed that automatic lesion segmentation is a none-solved issue and therefore, quantifying lesion characteristics and the subsequent effect on the fracture risk using deep learning will remain challenging.

Funding sources

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of generative AI in scientific writing

None.

CRediT authorship contribution statement

Ali Ataei: Writing – review & editing, Writing – original draft, Visualization, Software, Formal analysis. **Florieke Eggermont:** Writing – review & editing, Supervision, Software, Methodology, Formal analysis, Conceptualization. **Nico Verdonschot:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Nikolas Lessmann:** Writing – review & editing, Supervision, Software, Methodology. **Esther Tanck:** Writing – review & editing, Supervision, Methodology, Conceptualization.

Declaration of competing interest

None.

Data availability

Data will be made available on request.

References

[1] F. Macedo, et al., Bone metastases: an overview, Oncol. Rev. 11 (1) (2017) 321.
[2] J.R. Gralow, et al., NCCN task force report: bone health in cancer care, J. Natl. Compr. Cancer Netw. 7 (Suppl_3) (2009) (p. S-1-S-32).
[3] A. Ataei, et al., Evaluation of inter- and intra-operator reliability of manual segmentation of femoral metastatic lesions, Int. J. Comput. Assist. Radiol. Surg. 16 (2021) 1841–1849.
[4] A.F. Chambers, et al., Critical steps in hematogenous metastasis: an overview, Surg. Oncol. Clin. N. Am. 10 (2) (2001) 243–255.
[5] D. Axelrod, A.M. Gazendam, M. Ghert, The surgical management of proximal femoral metastases: a narrative review, Curr. Oncol. 28 (5) (2021) 3748–3757.
[6] A. Angelini, et al., Treatment of pathologic fractures of the proximal femur, Injury 49 (2018) S77–S83.
[7] F. Eggermont, et al., Can patient-specific finite element models better predict fractures in metastatic bone disease than experienced clinicians? Bone Joint Res. 7 (6) (2018) 430–439.
[8] F. Eggermont, et al., Patient-specific finite element computer models improve fracture risk assessments in cancer patients with femoral bone metastases compared to clinical guidelines, Bone 130 (2020), 115101.
[9] F. Eggermont, et al., A patient-specific fracture risk assessment tool for femoral bone metastases: using the bone strength (BOS) score in clinical practice, Cancers 14 (23) (2022) 5904.
[10] M.A. Stadelmann, et al., Conventional finite element models estimate the strength of metastatic human vertebrae despite alterations of the bone's tissue and structure, Bone 141 (2020), 115598.
[11] T.S. Kaneko, et al., Mechanical properties, density and quantitative CT scan data of trabecular bone with and without metastases, J. Biomech. 37 (4) (2004) 523–530.
[12] F. Milletari, N. Navab, S.-A. Ahmadi, V-net: fully convolutional neural networks for volumetric medical image segmentation, in: 2016 Fourth International Conference on 3D Vision (3DV), IEEE, 2016.
[13] T. Kooi, et al., Large scale deep learning for computer aided detection of mammographic lesions, Med. Image Anal. 35 (2017) 303–312.
[14] J. Wang, et al., A multi-resolution approach for spinal metastasis detection using deep Siamese neural networks, Comput. Biol. Med. 84 (2017) 137–146.
[15] J. Chmelik, et al., Deep convolutional neural network-based segmentation and classification of difficult to define metastatic spinal lesions in 3D CT data, Med. Image Anal. 49 (2018) 76–88.
[16] H.R. Roth, et al., Detection of sclerotic spine metastases via random aggregation of deep convolutional neural network classifications, in: Recent Advances in Computational Methods and Clinical Applications for Spine Imaging, Springer, 2015, pp. 3–12.
[17] A. Madabhushi, G. Lee, Image analysis and machine learning in digital pathology: challenges and opportunities, Med. Image Anal. 33 (2016) 170–175.
[18] F. Isensee, et al., nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation, Nat. Methods 18 (2) (2021) 203–211.
[19] D. Rolnick, et al., Deep learning is robust to massive label noise, arXiv (2017), https://doi.org/10.48550/arXiv.1705.10694 (arXiv preprint arXiv: 1705.10694v3).
[20] F. Eggermont, et al., Calibration with or without phantom for fracture risk prediction in cancer patients with femoral bone metastases using CT-based finite element models, PLoS One 14 (7) (2019), e0220564.
[21] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-assisted Intervention, Springer, 2015.
[22] P. Refaeilzadeh, L. Tang, H. Liu, Cross-validation, in: Encyclopedia of Database Systems 5, 2009, pp. 532–538.
[23] J.H. Keyak, et al., Predicting proximal femoral strength using structural engineering models, Clin. Orthop. Relat. Res. 437 (2005) 219–228.

[24] I. Fleps, E.F. Morgan, A review of CT-based fracture risk assessment with finite element modeling and machine learning, Curr. Osteoporos. Rep. 20 (5) (2022) 309–319.
[25] I.R. Paranavithana, et al., Systematic review of tumor segmentation strategies for bone metastases, Cancers 15 (6) (2023) 1750.
[26] Y. Fang, et al., The impact of training sample size on deep learning-based organ auto-segmentation for head-and-neck patients, Phys. Med. Biol. 66 (18) (2021), 185012.
[27] N. Orlando, et al., Effect of dataset size, image quality, and image type on deep learning-based automatic prostate segmentation in 3D ultrasound, Phys. Med. Biol. 67 (7) (2022), 074002.
[28] S. Leclerc, et al., Deep learning applied to Multi-Structure segmentation in 2D echocardiography: a preliminary investigation of the required database size, in: 2018 IEEE International Ultrasonics Symposium (IUS), IEEE, 2018.
[29] P.A. Narayana, et al., Deep-learning-based neural tissue segmentation of MRI in multiple sclerosis: effect of training set size, J. Magn. Reson. Imaging 51 (5) (2020) 1487–1496.
[30] H. Ning, et al., Choosing an appropriate training set size when using existing data to train neural networks for land cover segmentation, Ann. GIS 26 (4) (2020) 329–342.
[31] O. Ciga, A.L. Martel, Learning to segment images with classification labels, Med. Image Anal. 68 (2021), 101912.
[32] C.S. Rajapakse, et al., Influence of bone lesion location on femoral bone strength assessed by MRI-based finite-element modeling, Bone 122 (2019) 209–217.
[33] J.H. Keyak, et al., Predicting the strength of femoral shafts with and without metastatic lesions, Clin. Orthop. Relat. Res. 439 (2005) 161–170.
[34] T.A. Damron, K.A. Mann, Fracture risk assessment and clinical decision making for patients with metastatic bone disease, J. Orthop. Res. 38 (6) (2020) 1175–1190.
[35] L.C. Derikx, et al., The assessment of the risk of fracture in femora with metastatic lesions: comparing case-specific finite element analyses with predictions by clinical experts, J. Bone Joint Surg. Brit. 94 (8) (2012) 1135–1142.
[36] C.M. Whyne, S.S. Hu, J.C. Lotz, Burst fracture in the metastatically involved spine: development, validation, and parametric analysis of a three-dimensional poroelastic finite-element model, Spine 28 (7) (2003) 652–660.
[37] M. Costa, et al., Biomechanical assessment of vertebrae with lytic metastases with subject-specific finite element models, J. Mech. Behav. Biomed. Mater. 98 (2019) 268–290.