Advancing Patient-Specific Fracture Risk Assessment in Femoral Bone Metastases through Finite Element Modeling and Deep Learning



Ali Ataei

RADBOUD UNIVERSITY PRESS

Advancing Patient-Specific Fracture Risk Assessment in Femoral Bone Metastases through Finite Element Modeling and Deep Learning

Ali Ataei

Ali Ataei

Advancing Patient-Specific Fracture Risk Assessment in Femoral Bone Metastases through Finite Element Modeling and Deep Learning

Radboud Dissertation Series
ISSN: 2950-2772 (Online); 2950-2780 (Print)

Published by RADBOUD UNIVERSITY PRESS
Postbus 9100, 6500 HA Nijmegen, The Netherlands
www.radbouduniversitypress.nl

Design: ProefschriftOntwerp
Cover image: Ali Ataei
Printing: ProefschriftMaken/Pumbo

ISBN: 9789465153186
DOI: 10.54195/9789465153186
Free download at: https://doi.org/10.54195/9789465153186

© 2026 Ali Ataei

RADBOUD
UNIVERSITY
PRESS

This is an Open Access book published under the terms of Creative Commons Attribution-Noncommercial-NoDerivatives International license (CC BY-NC-ND 4.0). This license allows reusers to copy and distribute the material in any medium or format in unadapted form only, for noncommercial purposes only, and only so long as attribution is given to the creator, see http://creativecommons.org/licenses/by-nc-nd/4.0/.

Advancing Patient-Specific Fracture Risk Assessment in Femoral Bone Metastases through Finite Element Modeling and Deep Learning

Proefschrift ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.M. Sanders,
volgens besluit van het college voor promoties
in het openbaar te verdedigen op

maandag 23 november 2026
om 14.30 uur precies

door

Ali Ataei
geboren op 21 september 1987
te Tabriz, Iran

Promotoren:
Prof. dr. ir. N.J.J. Verdonschot
Prof. dr. E.J.M. Tanck

Copromotoren:
Dr. N. Lessmann
Dr. F.E. Eggermont

Manuscriptcommissie:
Prof. dr. I. Sechopoulos
Dr. P.G. Westhoff
Prof. dr. ir. C.I. Sánchez Gutiérrez (Universiteit van Amsterdam)

Advancing Patient-Specific Fracture Risk Assessment in Femoral Bone Metastases through Finite Element Modeling and Deep Learning

Dissertation to obtain the degree of doctor
from Radboud University Nijmegen
on the authority of the Rector Magnificus prof. dr. J.M. Sanders,
according to the decision of the Doctorate Board
to be defended in public on

Monday, November 23, 2026
at 2.30 pm

by

Ali Ataei
born on September 21, 1987
in Tabriz, Iran

Supervisors:
Prof. dr. ir. N.J.J. Verdonschot
Prof. dr. E.J.M. Tanck

Co-supervisors:
Dr. N. Lessmann
Dr. F.E. Eggermont

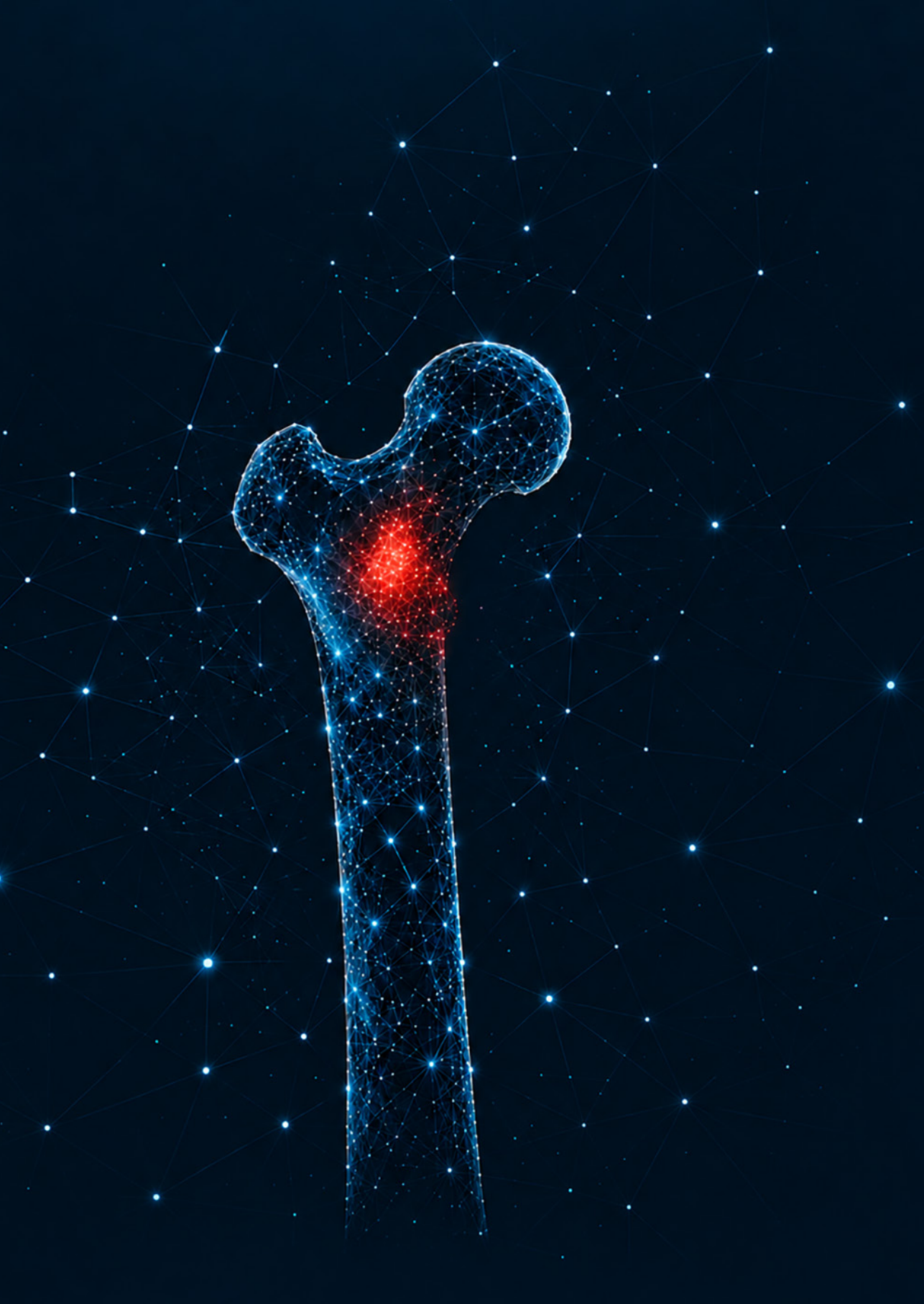
Manuscript Committee:
Prof. dr. I. Sechopoulos
Dr. P.G. Westhoff
Prof. dr. ir. C.I. Sánchez Gutiérrez (University of Amsterdam)

Dedicated to my parents

تقدیم به مادر و پدر عزیزم

TABLE OF CONTENTS

Chapter 1	General Introduction	11
Chapter 2	Evaluation of Inter- and Intra-operator Reliability of Manual Segmentation of Femoral Metastatic Lesions	25
Chapter 3	The effect of deep learning-based lesion segmentation on failure load calculations of metastatic femurs using finite element analysis	43
Chapter 4	Assessing the impact of density-adapted osteoblastic lesion in metastatic femur using finite element analysis and deep learning-based lesion segmentation	65
Chapter 5	The effect of variations in CT scan protocol on femoral finite element failure load assessment using phantomless calibration	85
Chapter 6	General Discussion	111
Chapter 7	Summary	129
Chapter 8	Samenvatting	137
Appendices	Research Data Management	147
	Radboudumc PhD Portfolio	148
	Curriculum Vitae	150
	Acknowledgements	151



Chapter 1

General Introduction

INTRODUCTION

Cancer and bone metastases

Bone is one of the most common sites for metastases originating from primary cancers originating from the breast, prostate, lung, kidney, and thyroid [1]. The skeletal system, with its extensive vascularization and continuous remodeling activity, provides an environment for the colonization of metastatic cancer cells. Among skeletal locations, the femur, spine, pelvis, and skull are frequently affected, with the femur being the most common site in the axial skeleton [2] (figure 1). The consequences of bone metastases can be severe, affecting not only skeletal integrity, but also the patient's overall quality of life. Pain, immobility, and pathological fractures are common, with the femur being the site of nearly 60% of all pathological fractures in long bones [3]. Survival rates for many cancer types have improved significantly over the past few decades. For example, five-year relative survival across all cancer types has risen from roughly 39 % in the 1960s to around 70 % in the 2020s in the U.S. [4]. As patients live longer, the need to maintain their quality of life becomes increasingly important and making proactive management of skeletal complications such as femoral metastases more essential than ever.



Figure 1. Frontal view of metastatic bone lesions in the proximal femur and diaphysis. The femur is one of the most common sites for bone metastases, which can severely compromise structural integrity and lead to pain, immobility, and pathological fractures.

Types of bone metastases

The clinical presentation of bone metastases can differ considerably. They are typically classified into three categories based on radiographic appearance: osteolytic, osteoblastic, and

mixed [1] (Figure 2). Osteolytic lesions involve increased bone resorption, leading to loss of structural integrity. In contrast, osteoblastic metastases stimulate the formation of immature and disorganized bone that, despite its increased density on imaging, lacks normal mechanical strength [5]. Mixed lesions contain elements of both, making their biomechanical assessment more complex.

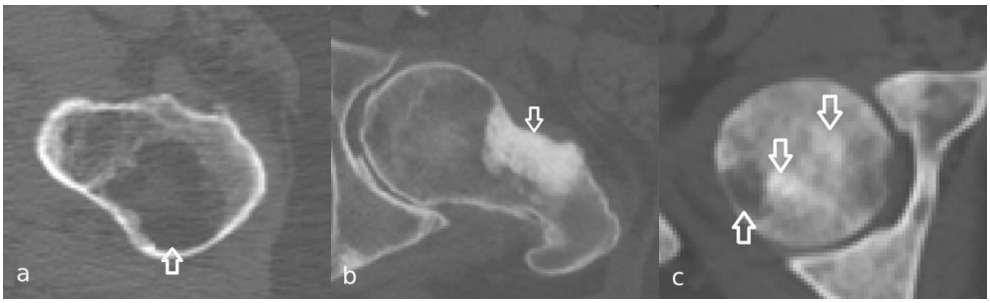


Figure 2. Cross-sectional CT scans of the proximal femur with bone metastases. Arrows indicate (a) an osteolytic lesion, (b) an osteoblastic lesion, and (c) a mixed lesion containing both osteolytic and osteoblastic features. Reproduced from [Ataei et al., 2021], with permission.

Clinical management of bone metastases

Bone metastases can cause several complications affecting patient's daily life and overall condition. Pain is a common early symptom of bone metastases [6]. As the lesions grow, they weaken the bone structure, making it more likely to fracture, even during normal daily activities [6-8]. Pathological fractures are one of the most serious complications in cancer patients with bone metastases, especially when the femur is affected. They often result in a sudden decrease in the patient's condition, leading to reduced mobility, difficulty with daily activities, and a lower quality of life [6]. These fractures usually require hospitalization, surgery, and a long period of recovery. They can also delay cancer treatment and are linked to a higher risk of death [9].

Given these severe clinical consequences, it is important to estimate the fracture risk accurately. Conservative treatment options such as radiotherapy, or pain management are generally preferred when the fracture risk is deemed low. In contrast, patients with a high fracture risk are considered for preventive fixation [10]. Surgical intervention prior to a fracture is associated with better outcomes compared to surgery performed post-fracture [11]. Fracture risk plays an important role in choosing the most suitable treatment, along with other clinical factors.

Traditional methods of fracture risk assessment

Traditionally, fracture risk in metastatic femoral bone disease has been assessed using plain radiography and tools such as Mirels' scoring system [12]. This system uses lesion site, size, radiographic appearance, and pain level to calculate a risk score. A score of nine or higher leads to recommendations for surgical intervention [12]. However, multiple studies have shown that Mirels' scoring system lacks specificity, leading to unnecessary surgeries in a significant proportion of patients [10, 13-15].

Axial cortical involvement, defined as 30 mm or more of continuous lesion involvement along the femoral cortex, has been proposed as an alternative criterion and has shown improved specificity [10, 13]. Nonetheless, both methods rely on subjective clinical judgment and qualitative imaging interpretation and fail to consider key patient-specific factors such as bone geometry, mechanical loading, and general bone quality.

Given the limitations of these methods, clinicians are often faced with uncertainty, potentially resulting in either overtreatment that exposes patients to unnecessary surgical risk and healthcare costs or undertreatment, which can lead to fractures and worsened prognosis [16]. For example, applying a Mirels' score threshold of higher than nine has been associated with low specificity (45%), leading to unnecessary surgeries in up to two-thirds of patients who would not have fractured, with positive predictive values (PPV) between 10% and 32% for femoral lesions [13-15]. In contrast, axial cortical involvement of ≥30 mm has shown improved specificity (50%) and high negative predictive value (96%). However, it still results in overtreatment in approximately 80% of patients classified as high risk (PPV=20%) [13].

Calculating bone strength using Finite Element (FE) modeling

FE modeling has become a valuable and objective tool for predicting fracture risk in femora affected by bone metastases [8, 16-20], as well as osteoporotic bones [21, 22]. Several research groups have contributed to this field. For example Keyak et al. demonstrated early applications of FE modeling to estimate femoral strength under various lesion conditions [19]. In the context of osteoporosis, studies by Schileo et al. and Orwoll et al. applied patient-specific FE models to assess fracture risk and improved clinical predictions beyond standard bone mineral density measurements [21, 22]. This technique starts with segmentation of the femur from patient-specific CT scan data to obtain its geometry, after which the scan data are calibrated using either a physical phantom or a phantomless method to convert Hounsfield units (HUs) into bone mineral density (BMD). Mechanical properties are assigned to the model elements based on local density values, and the femur is subjected to virtual loading under physiological conditions until failure. In our research group, this estimated failure load is normalized to the patient's body weight to calculate the BOne Strength (BOS) score, which provides an objective measure of fracture risk to support clinical decision-making [18] (Figure 3).

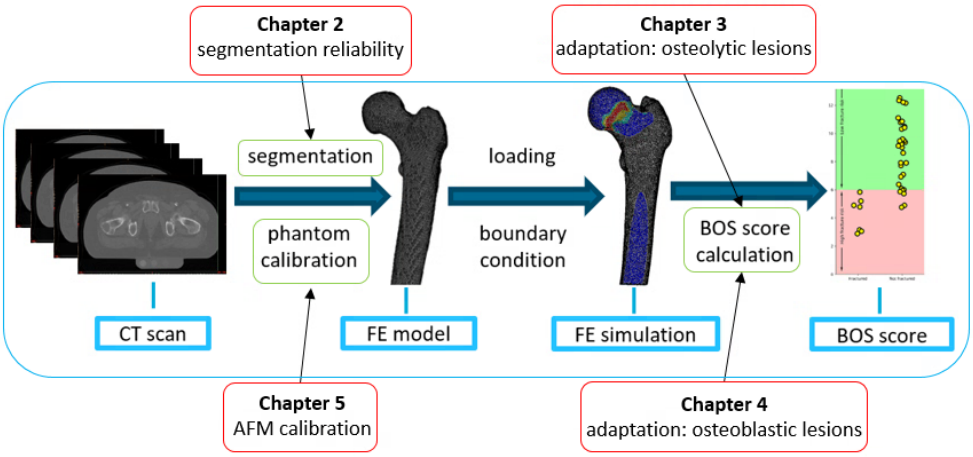


Figure 3. Overview of the finite element simulation flow for patient-specific fracture risk assessment.The process includes segmentation of the femur and calibration of Hounsfield units (using either phantom or air–fat–muscle (AFM) method) to generate an FE model, which is then subjected to physiological loading and boundary conditions in an FE simulation. The resulting failure load is used to calculate the BOne Strength (BOS) score. Each chapter of the thesis addresses a specific aspect of this workflow.

The BOS score has demonstrated higher sensitivity and specificity than traditional clinical scoring methods used for osteolytic femoral metastases, achieving 100% sensitivity and 74% specificity compared to 86% and 42% with conventional criteria, respectively. It also showed a higher positive predictive value (39% vs 19%) and negative predictive value (100% vs 95%) in this comparison. [18]. It classifies patients with osteolytic lesions into low or high fracture risk categories, helping clinicians and patients choose the most appropriate treatment. While its clinical use is expected to reduce unnecessary surgeries, current evidence mainly shows that the BOS score provides objective insight into fracture risk and facilitates shared decision-making between physicians and patients [16, 18]. Patient-specific FE modeling in osteolytic metastases has been shown to enhance the accuracy of fracture risk assessments compared to traditional clinical guidelines, allowing for more appropriate treatment decisions and reducing the likelihood of both overtreatment and undertreatment [16]. Moreover, the BOS score has been reported to improve clinical certainty and serve as a valuable tool in multidisciplinary discussions, contributing to more informed and collaborative treatment planning [18]. Thus, for osteolytic lesions the BOS score already provides a reasonably reliable basis for clinical decision-making, whereas the applicability to osteoblastic and mixed lesions remains less certain.

Limitations of the current Finite Element workflow

Despite the aforementioned advantages, FE modeling has several limitations. In standard practice, a calibration phantom is scanned alongside the patient to convert HUs into BMD values [23, 24]. Although this method is considered accurate, it is costly, impractical for routine use, and cannot be applied to existing scans that were acquired without a calibration phantom. To address these limitations, phantomless calibration methods have been developed [25-27]. An example of a phantomless method is the air-fat-muscle (AFM) approach, which uses the HU values of tissues like air, fat, and muscle as internal references to calibrate CT scans [25]. However, HU values are influenced by scanner type and acquisition protocols, which may introduce variability in phantomless calibration outcomes and limit their transferability across centers.

Another key limitation is the assumption that bone mechanical behavior can be predicted solely from BMD. While this assumption is reasonable for healthy or osteolytic bone, it fails in osteoblastic lesions [18]. These lesions often appear very dense on CT scans (Figure 2b) but are composed of disorganized, weak bone [28]. As a result, osteoblastic bone is modeled as being stiffer and stronger than its actual mechanical properties. This inconsistency limits the applicability of FE models for osteoblastic or mixed lesions and may result in underestimation of the fracture risk. To address this discrepancy, lesions must first be identified (i.e. segmented) in the CT scan so that their mechanical properties can be (manually) adjusted in the model, allowing for a more realistic simulation of fracture behavior.

Segmentation of the metastatic lesions, which can be performed manually or using automatic methods, is a challenging but essential preparatory step for assigning realistic material properties in the FE models. Manual segmentation, the current clinical standard, is time-consuming and may be subject to operator variability. Inter- and intra-operator differences can be considerable, especially in complex or diffuse lesions, which raises concerns about the objectivity of the strength calculations and highlights the need to systematically assess segmentation reliability.

Artificial intelligent (AI) solution for lesion segmentation:

The growing use of deep learning and artificial intelligence in medical imaging has provided new possibilities for improving the segmentation of metastatic bone lesions. Neural networks such as nnU-Net and its variants have shown robust performance and adaptability across a wide range of medical imaging tasks without the need for manual hyperparameter tuning [29-33]. nnU-Net has been successfully applied to segment femoral tumors [30], spinal metastases [32], and pelvic bone lesions [33], consistently achieving high Dice coefficients (typically>0.8), approaching inter-observer agreement levels seen between experienced radiologists [34].

These studies demonstrate that models such as nnU-Net can learn complex imaging patterns by training on large, annotated datasets, enabling them to distinguish between healthy and pathological tissue. This is particularly valuable in skeletal oncology, where metastases present with heterogeneous appearance (osteolytic, osteoblastic, and mixed lesions) that may be difficult to segment manually. Beyond lesion-level segmentation, deep learning has also been used for automatic segmentation of entire bones from CT or MRI images, facilitating the generation of patient-specific anatomical models for further computational processes such as FE modeling or surgical planning [29].

Recent studies have shown that integrating deep learning-based segmentations into FE methods can enhance the accuracy of metastasized bone modeling. A study by Rachmil et al. reported that automated segmentation of osteolytic lesions resulted in higher principal strain estimates in affected areas, closely matching with results from manual segmentation and models without lesion segmentation [35]. However, they did not assess how automatic segmentation affects predicted failure load or clinical fracture risk assessment.

Beyond the context of FE modeling, automatic segmentation is increasingly used in other clinical workflows. In radiotherapy planning, deep learning models have become standard for auto-contouring of tumors and organs-at-risk (OARs), improving consistency and dramatically reducing planning time [36]. In diagnostic radiology, convolutional neural network (CNN)-based tools have been shown to augment expert performance. For example, a study found that integrating a deep learning model into radiologists workflow significantly improved the detection of bone metastases on CT while halving interpretation time [31]. Moreover, general segmentation models such as Segment Anything Model (SAM) are now being explored for rapid initialization of medical image segmentation tasks, broadening the scalability of AI assistance across modalities [34]. In addition, automatic segmentation can facilitate advanced analyses through the extraction of radiomic features (quantitative descriptors of lesion phenotype, e.g., shape, intensity, and texture patterns derived from medical images) and pathomic features (quantitative descriptors from digitalized histopathology slides), supporting diagnosis, prognosis, and treatment response prediction [34].

In summary, AI-driven segmentation is a promising tool in skeletal oncology workflows, including fracture risk modeling, treatment planning, and diagnostic imaging. By leveraging CNN-based tools for segmentation and integrating them with patient-specific biomechanical simulations, this approach has the potential to reduce segmentation time, improve consistency across operators, and ultimately support more reliable fracture risk assessment.

Research goals and thesis questions

This thesis aims to advance the clinical applicability of FE modeling for fracture risk assessment in femora with osteoblastic or mixed bone metastases. Therefore, the primary goal was to improve material property assignment by developing an automatic segmentation tool to segment metastatic lesions thereby enabling adjustment of their mechanical properties.

In addition, the thesis investigates the feasibility of phantomless CT scan calibration to replace conventional phantom-based methods, which are often impractical in clinical and retrospective settings. Specifically, the goal was to evaluate the AFM calibration method across a range of clinical CT scan protocols to determine its consistency and potential integration into the FE modeling workflow. Generating a robust AFM calibration method amongst different CT scanning settings would allow for an easier implementation of the BOS score in multiple treatment centers in the Netherlands and internationally, and would facilitate utilization of retrospective CT scanning data sets (i.e. not containing the standard calibration phantom) of patients with femoral metastases.

The central research questions addressed in this thesis are:

- How reliable is the inter- and intra-operator manual segmentation of femoral metastatic lesions?
- Can deep learning models, such as nnU-Net, reliably segment femoral bone metastases across lesion types and match the accuracy of manual methods in FE outcomes?
- How can lesion-specific material properties be integrated into FE models to improve the accuracy of fracture risk prediction, especially in osteoblastic metastases?
- How do variations in CT scan protocols affect femoral finite element failure load assessment when using air-fat-muscle calibration method?

This thesis contributes to the advancement of biomechanical methods, and the BOS score in particular, that enables personalized treatment planning and help minimize unnecessary interventions, with the overall goal of improving care for patients with femoral bone metastases.

THESIS OUTLINE

In **chapter 2**, the manual segmentation of femoral osteolytic, osteoblastic, and mixed metastatic lesions was investigated. Manual segmentation, which is currently the standard method, may lack consistency. In this chapter, the reliability of manual segmentation was evaluated by comparing inter- and intra-operator results.

In **chapter 3**, lesion-specific FE modeling of osteolytic lesions was investigated. Osteolytic lesions were automatically segmented using nnU-Net framework, and their bone density was set to zero to simulate mechanical weakening. The resulting models were compared with standard CT-based models to evaluate the impact on failure load and clinical risk classification using the BOS score.

In **chapter 4**, the variation in material properties of osteoblastic lesions on bone strength as predicted by the FE modeling workflow was investigated. Lesion-specific material properties were adjusted to reflect their mechanical weakness, and the impact on BOS scores was evaluated. Manual segmentations were compared with those generated using nnU-Net framework.

In **chapter 5**, the effect of phantomless CT scan calibration and FE model outcomes was examined. A phantomless calibration approach using air, fat, and muscle as internal reference tissues was evaluated. The main objective was to assess how variations in CT scan protocols may influence failure load predictions when using the AFM method, and to compare these results with those obtained through phantom-based calibration.

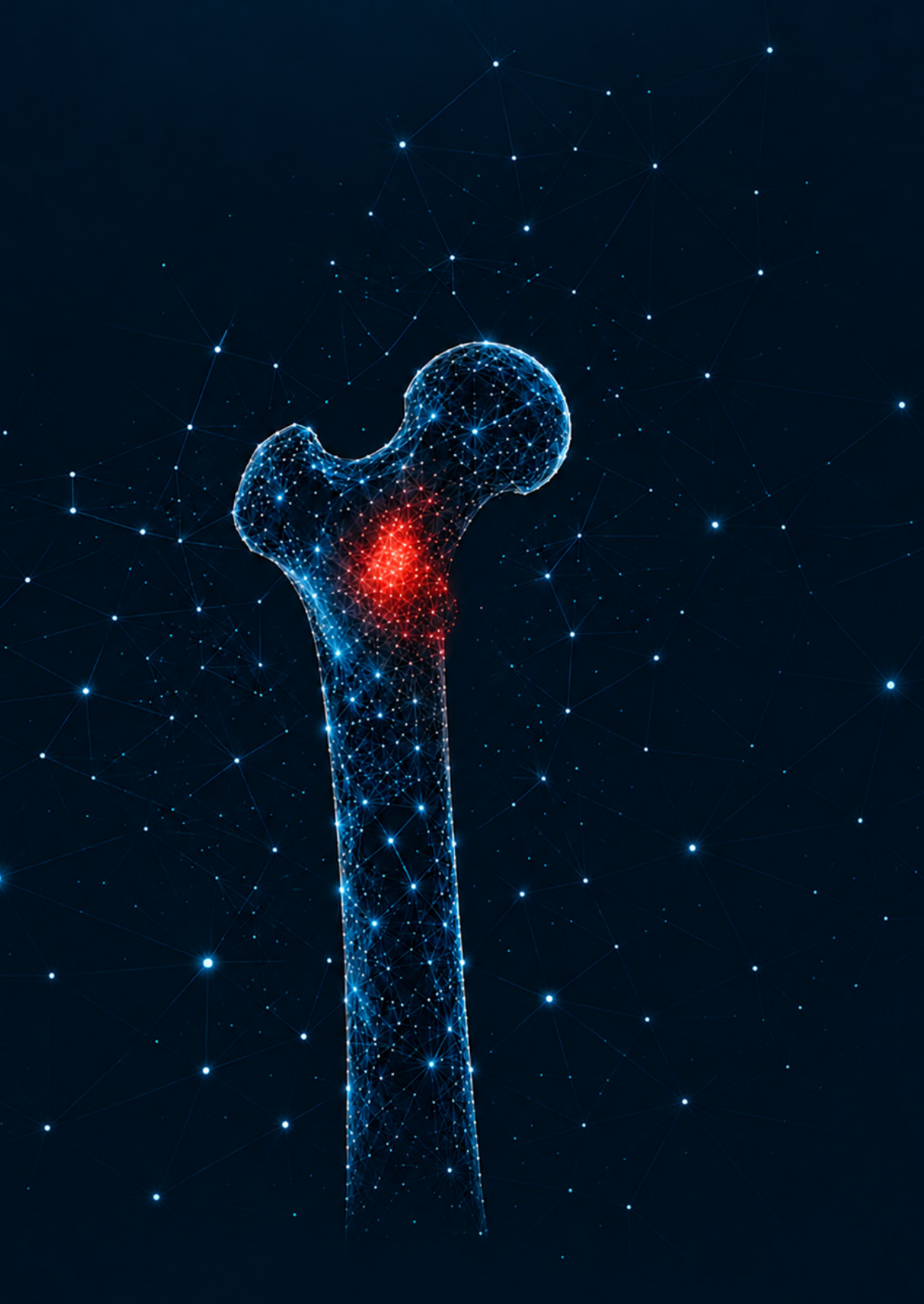
Finally, in **chapter 6** (Discussion) overall findings and contributions of this thesis were formulated. The effect of segmentation, calibration, and material properties assignment were discussed in relation to their potential to support more accurate fracture risk assessment in clinical settings and the perspectives for future research were formulated.

REFERENCES

1. Coleman, R., *Metastatic bone disease: clinical features, pathophysiology and treatment strategies.* Cancer treatment reviews, 2001. **27**(3): p. 165-176.
2. Axelrod, D., A.M. Gazendam, and M. Ghert, *The surgical management of proximal femoral metastases: a narrative review.* Current Oncology, 2021. **28**(5): p. 3748-3757.
3. Angelini, A., G. Trovarelli, A. Berizzi, E. Pala, A. Breda, M. Maraldi, and P. Ruggieri, *Treatment of pathologic fractures of the proximal femur.* Injury, 2018. **49**: p. S77-S83.
4. American Cancer Society, *Cancer Facts & Figures 2025*. 2025.
5. Roodman, G.D. and R. Silbermann, *Mechanisms of osteolytic and osteoblastic skeletal lesions.* BoneKEy reports, 2015. **4**: p. 753.
6. Johnson, S.K. and M.T. Knobf, *Surgical interventions for cancer patients with impending or actual pathologic fractures.* Orthopaedic Nursing, 2008. **27**(3): p. 160-171.
7. Tsuzuki, S., S.H. Park, M.R. Eber, C.M. Peters, and Y. Shiozawa, *Skeletal complications in cancer patients with bone metastases.* International Journal of Urology, 2016. **23**(10): p. 825-832.
8. Eggermont, F., L. Derikx, N. Verdonschot, I. Van Der Geest, M. De Jong, A. Snyers, Y. Van Der Linden, and E. Tanck, *Can patient-specific finite element models better predict fractures in metastatic bone disease than experienced clinicians?: Towards computational modelling in daily clinical practice.* Bone & joint research, 2018. **7**(6): p. 430-439.
9. Coleman, R.E., *Clinical features of metastatic bone disease and risk of skeletal morbidity.* Clinical cancer research, 2006. **12**(20): p. 6243s-6249s.
10. van der Linden, Y.M., H.M. Kroon, S.P. Dijkstra, J.J. Lok, E.M. Noordijk, J.W.H. Leer, and C.A. Marijnen, *Simple radiographic parameter predicts fracturing in metastatic femoral bone lesions: results from a randomised trial.* Radiotherapy and Oncology, 2003. **69**(1): p. 21-31.
11. DeHority, K., T. Craig, and T.A. Damron, *Prophylactic surgical treatment using CT-based rigidity analysis vs. after the fact fracture treatment of pathologic femoral lesions.* Annals of Joint, 2022. **7**: p. 12.
12. Mirels, H., *Metastatic Disease in Long Bones A Proposed Scoring System for Diagnosing Impending Pathologic Fractures.* Clinical Orthopaedics and Related Research (1976-2007), 1989. **249**: p. 256-264.
13. Van der Wal, C., F. Eggermont, M. Fiocco, H. Kroon, O. Ayu, A. Slot, A. Snyers, T. Rozema, N. Verdonschot, and P. Dijkstra, *Axial cortical involvement of metastatic lesions to identify impending femoral fractures; a clinical validation study.* Radiotherapy and oncology, 2020. **144**: p. 59-64.
14. Nazarian, A., V. Entezari, D. Zurakowski, N. Calderon, J.A. Hipp, J.C. Villa-Camacho, P.P. Lin, F.H. Cheung, A.J. Aboulafia, and R. Turcotte, *Treatment planning and fracture prediction in patients with skeletal metastasis with CT-based rigidity analysis.* Clinical Cancer Research, 2015. **21**(11): p. 2514-2519.
15. Altsitzioglou, P., S. Tsukamoto, C. Errani, Y. Tanaka, and A.F. Mavrogenis, *Predicting the Risk for Pathological Fracture in Bone Metastases.* Current Oncology, 2025. **32**(6): p. 309.
16. Eggermont, F., G. Van Der Wal, P. Westhoff, A. Laar, M. De Jong, T. Rozema, H.M. Kroon, O. Ayu, L. Derikx, and S. Dijkstra, *Patient-specific finite element computer models improve fracture risk assessments in cancer patients with femoral bone metastases compared to clinical guidelines.* Bone, 2020. **130**: p. 115101.

17. Derikx, L.C., J.B. van Aken, D. Janssen, A. Snyers, Y.M. van der Linden, N. Verdonschot, and E. Tanck, *The assessment of the risk of fracture in femora with metastatic lesions: comparing case-specific finite element analyses with predictions by clinical experts.* The Journal of Bone & Joint Surgery British Volume, 2012. **94**(8): p. 1135-1142.
18. Eggermont, F., Y. van der Linden, N. Verdonschot, E. Dierselhuis, S. Ligthert, T. Bitter, P. Westhoff, and E. Tanck, *A Patient-Specific Fracture Risk Assessment Tool for Femoral Bone Metastases: Using the Bone Strength (BOS) Score in Clinical Practice.* Cancers, 2022. **14**(23): p. 5904.
19. Keyak, J.H., T.S. Kaneko, H.B. Skinner, and B.H. Hoang, *The effect of simulated metastatic lytic lesions on proximal femoral strength.* Clinical Orthopaedics and Related Research®, 2007. **459**: p. 139-145.
20. Yosibash, Z., R.P. Mayo, G. Dahan, N. Trabelsi, G. Amir, and C. Milgrom, *Predicting the stiffness and strength of human femurs with real metastatic tumors.* Bone, 2014. **69**: p. 180-190.
21. Orwoll, E.S., L.M. Marshall, C.M. Nielson, S.R. Cummings, J. Lapidus, J.A. Cauley, K. Ensrud, N. Lane, P.R. Hoffmann, and D.L. Kopperdahl, *Finite element analysis of the proximal femur and hip fracture risk in older men.* Journal of bone and mineral research, 2009. **24**(3): p. 475-483.
22. Schileo, E., F. Taddei, A. Malandrino, L. Cristofolini, and M. Viceconti, *Subject-specific finite element models can accurately predict strain levels in long bones.* Journal of biomechanics, 2007. **40**(13): p. 2982-2989.
23. Reeves, J.M., N.K. Knowles, G.S. Athwal, and J.A. Johnson, *Methods for post hoc quantitative computed tomography bone density calibration: phantom-only and regression.* Journal of Biomechanical Engineering, 2018. **140**(9): p. 094501.
24. Sternheim, A., O. Giladi, Y. Gortzak, M. Drexler, M. Salai, N. Trabelsi, C. Milgrom, and Z. Yosibash, *Pathological fracture risk assessment in patients with femoral metastases using CT-based finite element methods. A retrospective clinical study.* Bone, 2018. **110**: p. 215-220.
25. Eggermont, F., N. Verdonschot, Y. Van Der Linden, and E. Tanck, *Calibration with or without phantom for fracture risk prediction in cancer patients with femoral bone metastases using CT-based finite element models.* PLoS One, 2019. **14**(7): p. e0220564.
26. Lee, D.C., P.F. Hoffmann, D.L. Kopperdahl, and T.M. Keaveny, *Phantomless calibration of CT scans for measurement of BMD and bone strength—inter-operator reanalysis precision.* Bone, 2017. **103**: p. 325-333.
27. Weaver, A.A., K.M. Beavers, R.C. Hightower, S.K. Lynch, A.N. Miller, and J.D. Stitzel, *Lumbar bone mineral density phantomless computed tomography measurements and correlation with age and fracture incidence.* Traffic injury prevention, 2015. **16**(sup2): p. S153-S160.
28. Roudier, M.P., C. Morrissey, L.D. True, C.S. Higano, R.L. Vessella, and S.M. Ott, *Histopathological assessment of prostate cancer bone osteoblastic metastases.* The Journal of urology, 2008. **180**(3): p. 1154-1160.
29. Isensee, F., P.F. Jaeger, S.A. Kohl, J. Petersen, and K.H. Maier-Hein, *nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation.* Nature methods, 2021. **18**(2): p. 203-211.
30. Rachmil, O., M. Artzi, M. Iluz, I. Druckmann, Z. Yosibash, and A. Sternheim, *Automatic segmentation of femoral tumors by nnU-net.* Clinical Biomechanics, 2024. **116**: p. 106265.
31. Noguchi, S., M. Nishio, R. Sakamoto, M. Yakami, K. Fujimoto, Y. Emoto, T. Kubo, Y. Iizuka, K. Nakagomi, and K. Miyasa, *Deep learning–based algorithm improved radiologists' performance in bone metastases detection on CT.* European Radiology, 2022. **32**(11): p. 7976-7987.

32. Kim, D.H., J. Seo, J.H. Lee, E.-T. Jeon, D. Jeong, H.D. Chae, E. Lee, J.H. Kang, Y.-H. Choi, and H.J. Kim, *Automated detection and segmentation of bone metastases on spine MRI using U-Net: a multicenter study.* Korean Journal of Radiology, 2024. **25**(4): p. 363.
33. Liu, X., C. Han, Y. Cui, T. Xie, X. Zhang, and X. Wang, *Detection and segmentation of pelvic bones metastases in MRI images for patients with prostate cancer based on deep learning. Front Oncol. 2021; 11: 773299*. 2021.
34. Rich, J.M., L.N. Bhardwaj, A. Shah, K. Gangal, M.S. Rapaka, A.A. Oberai, B.K. Fields, G.R. Matcuk Jr, and V.A. Duddalwar, *Deep learning image segmentation approaches for malignant bone lesions: a systematic review and meta-analysis.* Frontiers in Radiology, 2023. **3**: p. 1241651.
35. Rachmil, O., K. Myers, O. Merose, A. Sternheim, and Z. Yosibash, *The influence of femoral lytic tumors segmentation on autonomous finite element analysis.* Clinical Biomechanics, 2024. **112**: p. 106192.
36. Erdur, A.C., D. Rusche, D. Scholz, J. Kiechle, S. Fischer, O. Llorian-Salvador, J.A. Buchner, M.Q. Nguyen, L. Etzel, and J. Weidner, *Deep learning for autosegmentation for radiotherapy treatment planning: State-of-the-art and novel perspectives.* Strahlentherapie und Onkologie, 2025. **201**(3): p. 236-254.



Chapter 2

Evaluation of Inter- and Intra-operator Reliability of Manual Segmentation of Femoral Metastatic Lesions

Ali Ataei[a*], Florieke Eggermont[a], Milan Baars[a], Yvette van der Linden[b], Jacky de Rooy[c], Nico Verdonschot[a, d], Esther Tanck[a]

[a]Radboud university medical center, Radboud Institute for Health Sciences, Orthopaedic Research Lab, P.O. Box 9101, 6500 HB Nijmegen, The Netherlands

[b]Department of Clinical Oncology, Leiden University Medical Center, The Netherlands

[c]Department of Radiology, Radboud university medical center, Nijmegen, The Netherlands

[d]Laboratory for Biomechanical Engineering, University of Twente, Enschede, The Netherlands

Published in

International Journal of Computer Assisted Radiology and Surgery

ABSTRACT

Purpose:

Accurate identification of metastatic lesions is important for improvement of biomechanical models that calculate the fracture risk of metastatic bones. The aim of this study was therefore to assess the inter- and intra-operator reliability of manual segmentation of femoral metastatic lesions.

Methods:

CT scans of 54 metastatic femurs (19 osteolytic, 17 osteoblastic, and 18 mixed) were segmented two times by two operators. Dice coefficients (DCs) were calculated adopting the quantification that a DC>0.7 indicates good reliability.

Results:

Generally, rather poor inter- and intra-operator reliability of lesion segmentation were found. Inter-operator DCs were 0.54 (±0.28) and 0.50 (±0.32) for the first and second segmentation, respectively, whereas intra-operator DCs were 0.56 (±0.28) for operator I and 0.71 (±0.23) for operator II. Larger lesions scored significantly higher DCs in comparison with smaller lesions. Of the femurs with larger mean segmentation volumes, 83% and 93% were segmented with good inter- and intra-operator DCs (>0.7), respectively. There was no difference between the mean DCs of osteolytic, osteoblastic, and mixed lesions.

Conclusion:

Manual segmentation of femoral bone metastases is very challenging and resulted in unsatisfactory mean reliability values. There is a need for development of a segmentation protocol to reduce the inter- and intra-operator segmentation variation as the first step and use of computer-assisted segmentation tools as a second step as this study shows that manual segmentation of femoral metastatic lesions is highly challenging.

Keywords: Bone metastasis; Manual segmentation; Reliability; Femoral lesions

1. INTRODUCTION

Bone is one of the most invaded tissues by the metastases of breast, prostate, lung, kidney, and thyroid cancer [1]. Approximately half of the primary tumors of these cancers metastasize to bone, with the femur, hip, spine, and skull as the most frequent sites [2]. Bone metastases can have three different types of appearance: osteolytic, osteoblastic, and mixed (Fig. 1). In osteolytic and osteoblastic metastases, bone absorption and bone formation are increased, respectively. Newly formed bone of osteoblastic metastases is known to be immature and of poor quality [3]. Bone metastases are often painful and may increase the risk of pathological fracture [4]. These pathological fractures result in increased rates of morbidity and mortality [5].

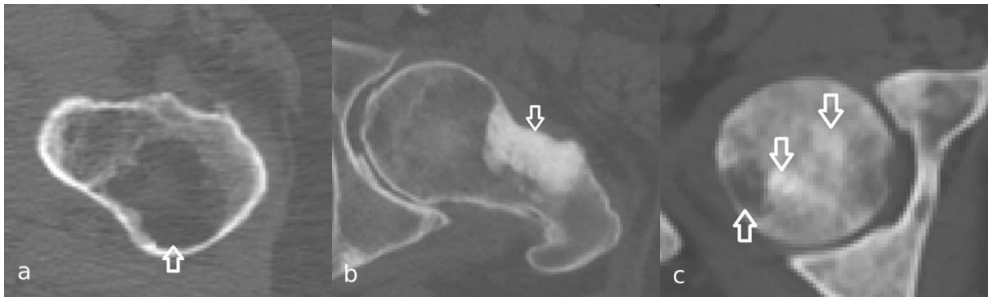


Fig. 1 Arrows indicate femoral osteolytic (a), osteoblastic (b), and mixed (c) bone metastases in axial view.

Patients with bone metastases are treated based on fracture risk: patients with a high fracture risk are considered for prophylactic surgery to prevent pathological fracture, whereas patients with a low fracture risk are treated with conservative treatment such as radiotherapy for pain management. Higher dose radiotherapy in order to induce remineralization is an option if the patient is not eligible for surgery [6].

Biomechanical finite element (FE) models can be used to evaluate metastatic bone fracture risk [7, 8]. In a prospective cohort study of our group, quantitative computed tomography (CT) based FE models were created and bone strengths of femurs with predominantly osteolytic metastatic lesions were calculated to predict whether or not a fracture would occur [7, 8]. FE predictions had a higher sensitivity compared to those from clinical assessments. However, currently, the FE models are not yet applicable to femurs affected with osteoblastic metastases. To determine the mechanical properties within the FE model, calibrated bone densities from CT scans are obtained and since osteoblastic metastases appear very dense on CT scans this results in strong mechanical properties of the metastatic lesions although they are weaker in reality. A potential solution is to assign more appropriate material properties to the osteoblastic

lesions to better simulate the mechanical behavior of the weakened tissue. For this purpose, exact segmentation of the metastatic lesions is important.

Additionally, segmentation of the metastatic lesions is important for radiotherapy planning. For this purpose, bone metastases are currently usually identified during visual inspection on CT scans and are segmented manually to determine exact radiation fields. Studies suggest that it is difficult to indicate exact edges of bone metastases, mainly in case of diffuse lesions [9-11]. Furthermore, studies investigated the inter-observer agreement of manual segmentation of bone metastases in the pelvis [12, 13], femoral head, and spine [12], as well as the scapula, humeral head, and ribs [13]. Their segmentations were done by a radiation oncologist, and relatively moderate agreement rates (0.46 and 0.61, respectively) were reported indicating that manual segmentation is challenging. However, the accuracy and repeatability of segmentation of femoral osteolytic, osteoblastic, and mixed bone metastases has not been studied before. Therefore the aim of our study was to assess the inter- and intra-operator reliability of manual segmentation of femoral metastatic lesions. This was assessed for osteolytic, osteoblastic and mixed lesions.

2. METHODS

To address the study aim, we used CT scans of patients with metastatic lesions which were segmented by two operators. Subsequently, the inter- and intra-operator reliability of the segmentations was assessed.

2.1. CT scan

From our existing patient cohort (Approved by Commissie Mensgebonden Onderzoek regio Arnhem-Nijmegen. Reference number 2013/305) [8], we included 54 CT scans of femurs with bone metastases (19 osteolytic, 17 osteoblastic and 18 mixed femurs) originating from multiple types of cancer such as prostate, breast, kidney, lung, esophagus, and plasma cell cancer (Table 1 appendix). All patients gave informed consent. CT scans were obtained from four radiotherapy institutes (Radboud university medical center, Leiden University Medical Center, Radiotherapeutic Institute Friesland Leeuwarden and Bernard Verbeeten Institute Tilburg). As these images came from a multi-center clinical study, different voxel sizes were found (voxel size ranged from 0.86x0.86x2.5 mm3 to 1.27x1.27x3 mm^3). Therefore, all 3-D lesion dimensions were displayed in cm^3.

2.2. Lesion segmentation

Two operators with background in biomedical engineering (operator I) and biomedical sciences (operator II), who were familiar with segmentation, were trained by an experienced radiologist to identify and segment metastatic lesions. Prior to the start of the segmentation, operators

practiced lesion segmentation on a trial metastatic lesion dataset for approximately 15 hours followed by a feedback session with the experienced radiologist. Subsequently, they segmented all lesions of the 54 femurs twice with a four-week time interval, using Mimics 14.0 and 20.0 (Materialise, Leuven, Belgium (Fig. 2)).

Observers segmented the lesions, i.e. the voxels around the bone cortex or the boundary of the lesion itself were selected more accurately than the larger areas that are selected when lesions are segmented by a radiologist or a radiation oncologist. During segmentation, the following rules were used. First, we always searched for abnormalities in the bone relative to proximal and distal slices and to the contralateral femur. Second, lesions with altered cancellous bone were included in the segmentation. Third, missing bone in osteolytic lesions, especially in the diaphysis, was segmented. For this we again compared the normal thickness of the bone relative to slices proximal and distal to the segmented slice and to the contralateral femur. Fourth, the sclerotic rim of osteolytic lesions was excluded from the segmentation. In case of mixed lesions, the sclerotic rim was included when it was very difficult to differentiate between an osteolytic and osteoblastic metastasis. Finally, for osteoblastic lesions and osteoblastic parts of mixed lesions, the segmentations were not limited to the normal boundary of the femoral bone, thus if the bone metastases grew out of the normal boundaries they were included in the segmentation.

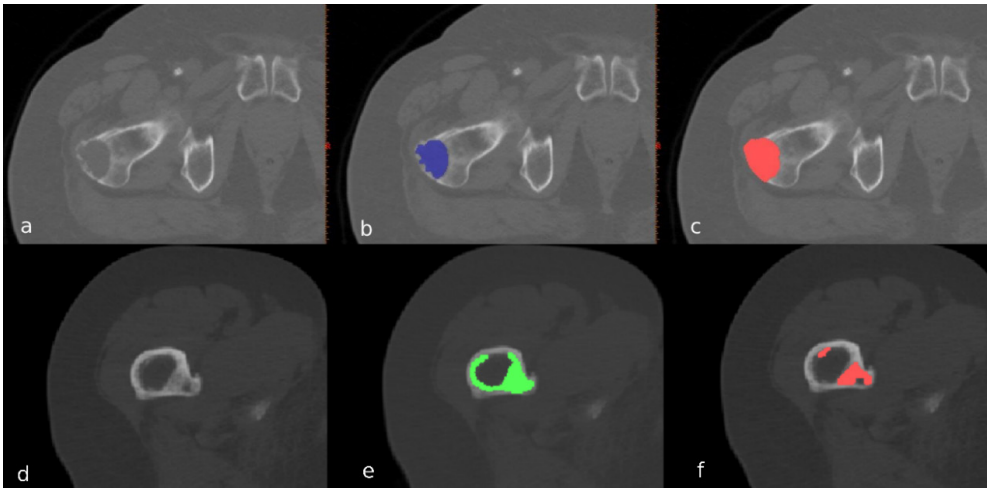


Fig. 2 Example of an osteolytic (upper half) and osteoblastic (lower half) lesion in the proximal femur, and the manual segmentations by operator I (b, e) and II (c, f) in axial view. Cortical voxels of the osteolytic lesion were included by operator II (c), but not by operator I (b). The operators did not agree on the number of osteoblastic lesions: operator I only segmented one lesion (e), whereas operator II segmented two separate lesions (f).

2.3. Inter- and intra-operator reliability

Segmentation similarities between and within operators were assessed using 3-D Dice coefficient (DC, formula 1). A and B represent two segmentation volumes:

$$dice\ coefficient = \frac{2\left|A \cap B\right|}{\left|A\right| + \left|B\right|} \qquad (1)$$

The DCs were calculated for each femur, including all lesions within that femur. This was done because operators did not necessarily agree on the number of lesions within one femur, causing some cases in which one operator segmented multiple smaller lesions, whereas the other operator segmented that as one larger lesion. We defined a DC>0.7 as a good reliability, in accordance with a statistical image segmentation study [14]. Since our data were not distributed normally, Kruskal–Wallis one-way analysis of variance was used to test the mean DC differences between osteolytic (n=19), osteoblastic (n=17), and mixed (n=18) lesions. P-values below 0.05 were considered statistically significant. Results of mean DC analysis are reported as mean ± standard deviation (SD).

In addition, segmentation differences, i.e. the non-overlapping segmentation volume between and within operators, were calculated (formula 2, A and B represent two segmentation volumes).

$$non-overlapping\ segmentation\ volume = \left|A\right| + \left|B\right| - 2\left|A \cap B\right| \quad (2)$$

2.4. Lesions size

To investigate whether the DC varies with the lesion size, mean segmentation volume of the first and second time (inter-operator) and operator I and II (intra-operator) were calculated per femur. A threshold of 60 cm^3 was chosen, based on the observation that DC above this threshold were generally larger than 0.7, below and beyond which we compared the DCs. Results of mean segmentation volume are reported as mean ± standard deviation (SD).

3. RESULTS

3.1. Data analyses

Only one out of four segmentation sets in this study (inter-operator I, inter-operator II, intra-operator I and intra-operator II) resulted in good mean DC (>0.7). Inter-operator DCs were 0.54 (±0.28) and 0.50 (±0.32) for the first and second segmentation, respectively, whereas intra-operator DCs were 0.56 (±0.28) for operator I and 0.71 (±0.23) for operator II. Table 1 summarizes the mean inter- and intra-operator DCs per lesion type and size. There was no statistically significant difference in the first and second segmentation mean DCs between osteolytic, osteoblastic, and mixed lesions (p = 0.86 and p = 0.41, respectively).

Table 1 Mean (±SD) inter- and intra-operator DCs and segmentation volume in cm^3 per lesion type.

		Inter-operator		Intra-operator	
		First segmentation	Second segmentation	Operator I	Operator II
Mean DC (±SD)	Osteolytic	0.54 (±0.26)	0.45 (±0.31)	0.56 (±0.26)	0.63 (±0.27)
	Osteoblastic	0.57 (±0.28)	0.60 (±0.27)	0.60 (±0.28)	0.78 (±0.12)
	Mixed	0.51 (±0.29)	0.46 (±0.34)	0.52 (±0.3)	0.74 (±0.23)
Mean Segmentation volume (±SD)	Osteolytic	14.48 (±13.83)	12.88 (±10.79)	14.44 (±13.65)	12.92 (±11.76)
	Osteoblastic	33.35 (±32.42)	37.12 (±40.56)	32.51 (±36.74)	37.95 (±37.77)
	Mixed	41.2 (±27.47)	35 (±28.39)	28.72 (±24.78)	47.48 (±35.82)

The mean inter-operator non-overlapping segmentation volumes of the first and second segmentations were 22.66 (±24.22) and 20.69 (±22.47) cm^3, respectively, whereas the mean intra-operator non-overlapping segmentation volumes were 14.53 (±12.74) and 13.94 (±16.51) cm^3 for operator I and II, respectively. The non-overlapping segmentation volume increased with increasing mean segmentation volume (Fig. 3).

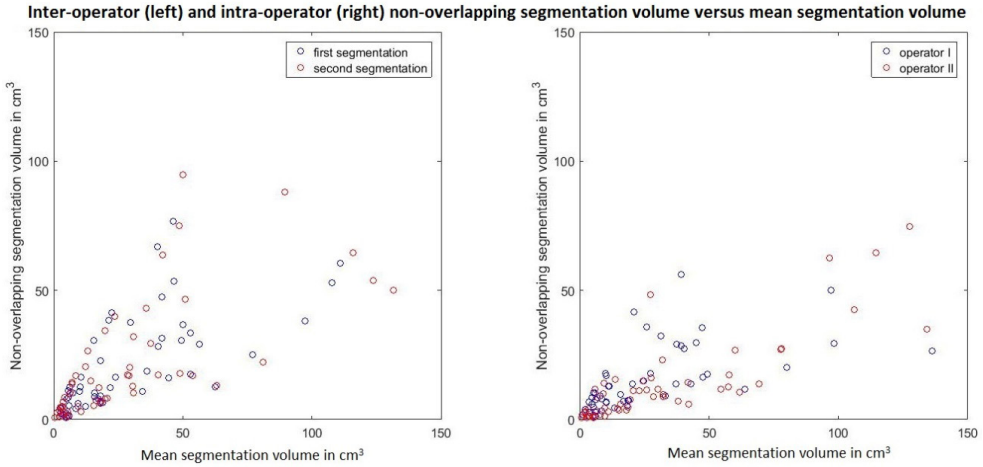


Fig. 3 Inter- and intra-operator non-overlapping segmentation volume versus mean segmentation volume. Each dot represents all segmented lesions of one femur. In the inter-operator figure (left), the blue dots represent the first segmentations and the red dots the second segmentations. In the intra-operator figure (right), the blue dots represents the segmentations by operator I and the red dots represent the segmentations by operator II.

3.2. Lesion size

For operator I, the mean segmentation volumes were 22.75 (±28.39) and 22.03 (±27.66) cm^3 for the first and second segmentation, respectively, whereas these were 30.91 (±32.24) and 33.74 (±37.32) cm^3 for the first and second segmentation of operator II, respectively. In general, the inter- and intra-operator DCs were higher for femurs with larger lesion segmentation volumes (Fig. 4). Accordingly, for femurs with larger lesion segmentation volumes (>60 cm^3), mean inter- and intra-operator DCs were good (>0.7), whereas for femurs with smaller lesion segmentation volumes, the DCs were lower (Table 2). Of the femurs with larger mean segmentation volumes, 83% and 93% were segmented with good inter- and intra-operator DCs (>0.7), respectively, whereas only 39% and 50% of the femurs with smaller mean segmentation volumes scored good inter- and intra-operator DCs, respectively (Fig. 5).

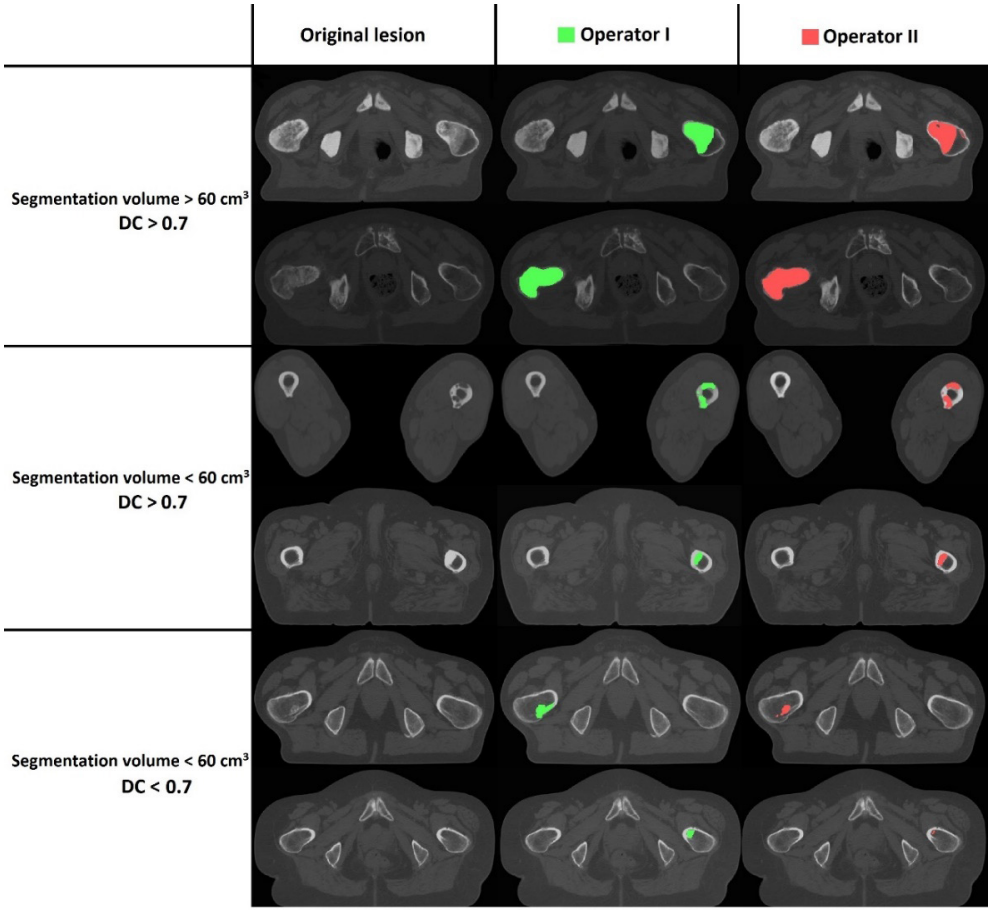


Fig. 4 Examples of segmentation comparisons between operator I and II for different lesion volumes and DCs above or below 0.7. Each row shows the axial view of a metastatic lesion either in the proximal part or in the diaphysis.

Table 2 Mean Inter- and intra-operator DCs per lesion size.

		Inter-operator	Intra-operator
Mean DC (±SD)	>60 cm^3	0.74 (±0.16)	0.81 (±0.07)
	<60 cm^3	0.49 (±0.3)	0.61 (±0.28)

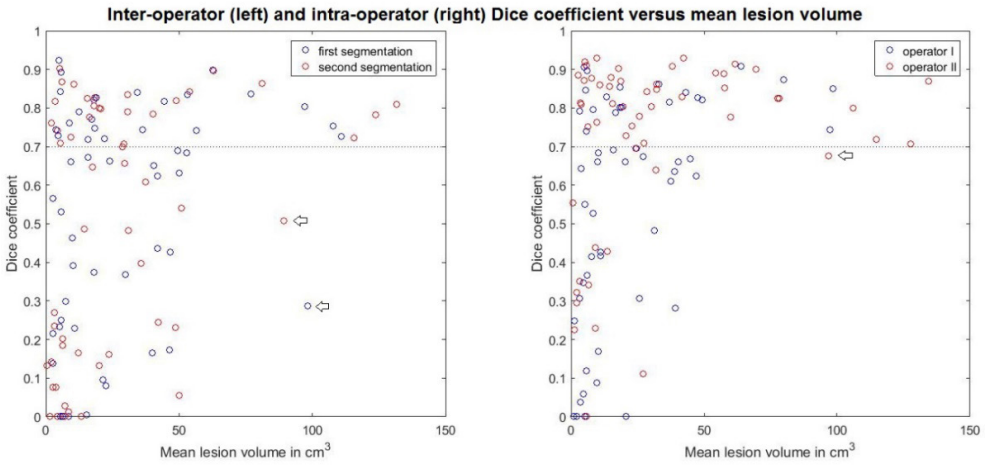


Fig. 5 Inter- and intra-operator DC versus mean segmentation volume. Note that in the inter-operator figure (left) mean segmentation volume indicates the mean over segmentation volume of operator I and II.

The blue dots represent the first measurements and the red dots represent the second measurements. The arrows indicate two femurs which scored DCs below 0.7 although they are in the group of larger lesions (>60 cm^3). In the intra-operator figure (right), mean segmentation volume indicates the mean over segmentation volumes of the first and second segmentations of each individual operator. The blue dots represent operator I and the red dots represent operator II. The arrow indicates a femur that scored DC below 0.7, although it was in the group of larger lesions (>60 cm^3).

4. DISCUSSION

In the current study, we aimed to evaluate the inter- and intra-operator reliability of manual segmentation of osteolytic, osteoblastic and mixed femoral metastatic lesions. For this purpose, metastatic lesions of 54 femurs were segmented twice by two operators. Since in most cases it was difficult to detect the bone metastases, only one out of four segmentation sets (inter-operator I, inter-operator II, intra-operator I and intra-operator II) resulted in good mean DC (>0.7). There was no statistically significant difference in the mean inter- and intra-operator DCs between osteolytic, osteoblastic, and mixed lesions.

The results of the segmentations showed that despite the same learning experience and instructions, segmentation remains a subjective procedure. The two operators in our study, who had no clinical experience, were trained by an expert radiologist. However, in our opinion they were adequately trained for this specific task and there are no indications that better segmentation agreements are achieved if they are done by domain experts [15]. Hence, the lesions in this

study were not segmented by an experienced radiologist. Additionally, when demonstrating a subset consisting of four segmentations of the same femora to the clinical experts, they were deemed adequate. This also proves the extreme difficulty of segmenting these types of lesions in a reproducible and robust manner, also for radiologists with clinical experience.

The results also showed that both non-overlapping segmentation volume and DC varied with the size of the lesion. Non-overlapping segmentation volume increased with an increase in lesion size. The results also convincingly showed that a high percentage (83% and 93%) of the larger lesions scored good inter- and intra-operator DC.

For most of the lesions, the segmentation volume differences originated from unclear boundaries which made it difficult for the operators to select the correct metastatic voxels during their visual inspections. In agreement with our findings, moderate mean agreement of 0.46 (range 0.15-0.75) and 0.61 (range 0.31-0.81) were reported by Gerlich et al. [12] and Raman et al. [13], respectively. Moreover, a study by Hoyte et al. [16] reported that the manual segmentation of the female pelvis floor organs was adequate for locating the organs but poor for segmenting structural boundaries. This is one of the most important points in differentiating between the ability of humans to recognize abnormal patterns and automated computer-assisted systems.

Similar studies in the past reported different reliability rates of manual segmentations of tissue abnormalities or whole target organs. For instance, Fiez et al. [17] reported a good agreement between their observers investigating the inter- and intra-operator reliability on segmentation of brain damage lesions, while Tingelhoff et al. [18] studied the replicability of the manual segmentation of sinuses and concluded that this is a time-consuming procedure which is not replicable. Accordingly, they suggested using automatic segmentation approaches.

This study has some limitations. The first limitation concerns our outcome measure. In contrast with the studies that use boundary-based measurements such as the Hausdorff distance to investigate the segmentation similarity, we used the volume-based Sorensen-Dice coefficient as an outcome measure since our operators did not necessarily agree on the number of lesions within a femur (Fig. 2).

As a second limitation, it should be noted that we obtained four lesion segmentation sets for each of the 54 femurs, and relatively large differences were typically found, for which it was impossible to know which segmentation was closest to quantify the true lesion in the bone. Hence, there was no ground truth. The absence of a ground truth could be addressed by utilizing cadaver bones in which lesions are drilled simulating osteolytic lesions. The empty holes will show maximal contrast and it is expected to have the maximum inter- and intra-operator reliability as there would be a clear CT intensity difference of the lesion with respect

to the non-lesion surrounding tissue. However, the translation to true clinical lesions would be difficult to make as in reality there will be a mixture of osteolytic and osteoblastic lesions. For this reason, such a cadaver assessment was not implemented in this study.

5. CONCLUSION

In conclusion, manual segmentation of femoral bone metastases is very challenging and resulted in unsatisfactory mean reliability values (DC between 0.4 and 0.7). Hence, development of a segmentation protocol to reduce the inter- and intra-operator segmentation variation as the first step, and development and use of a computer-assisted segmentation method as the second step, are warranted. Automatic segmentation methods based on deep learning algorithms may assist clinicians in their attempts to adequately segment lesions in the metastasized femurs and technicians to calculate bone strength of femurs with osteoblastic lesions.

6. REFERENCES

1. Coleman RE. Skeletal complications of malignancy. *Cancer: Interdisciplinary International Journal of the American Cancer Society*. 1997;80(S8):1588-94.
2. Johnson SK, Knobf MT. Surgical interventions for cancer patients with impending or actual pathologic fractures. *Orthopedic nursing*. 2008;27(3):160-71; quiz 72-3.
3. David Roodman G, Silbermann R. Mechanisms of osteolytic and osteoblastic skeletal lesions. *BoneKEy Rep*. 2015;4.
4. Tsuzuki S, Park SH, Eber MR, Peters CM, Shiozawa Y. Skeletal complications in cancer patients with bone metastases. *International Journal of Urology*. 2016;23(10):825-32.
5. Panula J, Pihlajamäki H, Mattila VM, Jaatinen P, Vahlberg T, Aarnio P, Kivelä S-L. Mortality and cause of death in hip fracture patients aged 65 or older - a population-based study. *BMC Musculoskeletal Disorders*. 2011;12(1):105.
6. van der Linden YM, Kroon HM, Dijkstra SPDS, Lok JJ, Noordijk EM, Leer JWH, Marijnen CAM. Simple radiographic parameter predicts fracturing in metastatic femoral bone lesions: results from a randomised trial. *Radiotherapy and Oncology*. 2003;69(1):21-31.
7. Eggermont F, Derikx LC, Verdonschot N, Geest ICMvd, Jong MAAd, Snyers A, Linden YMvd, Tanck E. Can patient-specific finite element models better predict fractures in metastatic bone disease than experienced clinicians? *Bone & Joint Research*. 2018;7(6):430-9.
8. Eggermont F, van der Wal G, Westhoff P, Laar A, de Jong M, Rozema T, Kroon HM, Ayu O, Derikx L, Dijkstra S, Verdonschot N, van der Linden Y, Tanck E. Patient-specific finite element computer models improve fracture risk assessments in cancer patients with femoral bone metastases compared to clinical guidelines. *Bone*. 2020;130:115101.
9. Dijkstra PDS, Oudkerk M, Wiggers T. Prediction of pathological subtrochanteric fractures due to metastatic lesions. *Arch Orthop Traum Su*. 1997;116(4):221-4.
10. Keene JS, Sellinger DS, Mcbeath AA, Engber WD. Metastatic Breast-Cancer in the Femur - a Search for the Lesion at Risk of Fracture. *Clin Orthop Relat R*. 1986(203):282-8.
11. Tatar Z, Soubrier M, Dillies AF, Verrelle P, Boisgard S, Lapeyre M. Assessment of the risk factors for impending fractures following radiotherapy for long bone metastases using CT scan-based virtual simulation: a retrospective study. *Radiat Oncol*. 2014;9.
12. Gerlich AS, van der Velden JM, Kotte AN, Tseng C-L, Fanetti G, Eppinga WS, Kasperts N, Intven MP, Pameijer FA, Philippens ME. Inter-observer agreement in GTV delineation of bone metastases on CT and impact of MR imaging: A multicenter study. *Radiotherapy and Oncology*. 2018;126(3):534-40.
13. Raman S, Chin L, Erler D, Atenafu EG, Cheung P, Chu W, Chung H, Loblaw A, Poon I, Rubenstein J. Impact of magnetic resonance imaging on gross tumor volume delineation in non-spine bony metastasis treated with stereotactic body radiation therapy. *International Journal of Radiation Oncology* Biology* Physics*. 2018;102(4):735-43. e1.
14. Zou KH, Warfield SK, Bharatha A, Tempany CM, Kaus MR, Haker SJ, Wells WM, 3rd, Jolesz FA, Kikinis R. Statistical validation of image segmentation quality based on a spatial overlap index. *Academic radiology*. 2004;11(2):178-89.

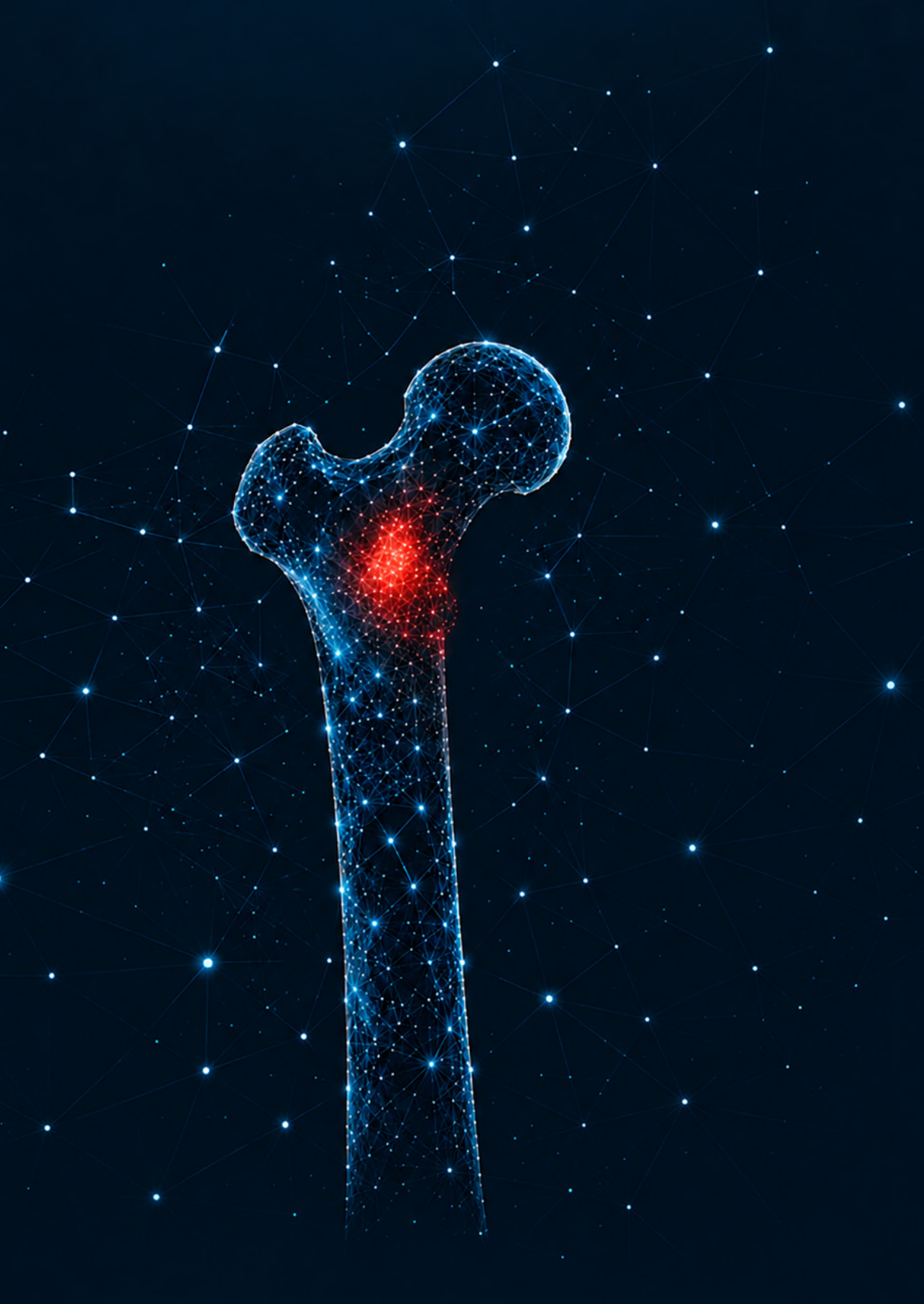
15. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, Van Der Laak JA, Van Ginneken B, Sánchez CI. A survey on deep learning in medical image analysis. *Medical image analysis*. 2017;42:60-88.
16. Hoyte L, Ye W, Brubaker L, Fielding JR, Lockhart ME, Heilbrun ME, Brown MB, Warfield SK. Segmentations of MRI images of the female pelvic floor: a study of inter- and intra-reader reliability. *Journal of magnetic resonance imaging : JMRI*. 2011;33(3):684-91.
17. Fiez JA, Damasio H, Grabowski TJ. Lesion segmentation and manual warping to a reference brain: intra- and interobserver reliability. *Hum Brain Mapp*. 2000;9(4):192-211.
18. Tingelhoff K, Eichhorn KWG, Wagner I, Kunkel ME, Moral AI, Rilk ME, Wahl FM, Bootz F. Analysis of manual segmentation in paranasal CT images. *European Archives of Oto-Rhino-Laryngology*. 2008;265(9):1061-70.

APPENDIX

Table 1 information of all 54 patients included in the study.

Patient ID	Sex (m=male; f=female)	Age (years)	Primary cancer type	Radiotherapy treatment
PT-01	m	55	Prostate	1x8 Gy
PT-02	m	59	Kidney	2x8 Gy
PT-03	m	82	Lung	5x4 Gy
PT-04	m	74	Prostate	1x8 Gy
PT-05	m	68	Multiple Myeloma	1x8 Gy
PT-06	f	61	Lung	1x8 Gy
PT-07	f	50	Breast	1x8 Gy
PT-08	m	73	Prostate	1x8 Gy
PT-09	m	74	Non-Hodgkin Lymphoma	10x3 Gy
PT-10	f	66	Breast	1x8 Gy
PT-11	f	60	Lung	After surgery 5x4 Gy
PT-12	m	70	Prostate	5x4 Gy
PT-13	f	68	Breast	1x8 Gy
PT-14	m	65	Prostate	1x8 Gy
PT-15	m	72	Prostate	1x8 Gy
PT-16	m	87	Prostate	1x8 Gy
PT-17	m	75	Prostate	1x8 Gy
PT-18	m	72	Prostate	1x8 Gy
PT-19	m	75	Prostate	5x4 Gy
PT-20	m	95	Prostate	No RT
PT-21	f	41	Breast	1x8 Gy
PT-22	f	67	Lung	1x8 Gy
PT-23	f	74	Multiple Myeloma	1x8 Gy
PT-24	f	55	Kidney	5x4 Gy
PT-25	m	60	Multiple Myeloma	Not available
PT-26	m	80	Prostate	1x8 Gy
PT-27	m	70	Prostate	5x4 Gy
PT-28	m	80	Prostate	1x8 Gy
PT-29	m	82	Prostate	1x8 Gy
PT-30	f	72	Lung	1x8 Gy
PT-31	m	87	Prostate	Not available
PT-32	f	61	Breast	1x8 Gy
PT-33	f	69	Multiple Myeloma	5x4 Gy
PT-34	m	53	Lung	1x8 Gy
PT-35	m	66	Prostate	1x8 Gy

PT-36	m	77	Multiple Myeloma	5x4 Gy
PT-37	m	48	Lung	1x8 Gy
PT-38	m	55	Prostate	1x8 Gy
PT-39	f	51	Lung	1x8 Gy
PT-40	f	65	Lung	1x8 Gy
PT-41	f	67	Lung	1x8 Gy
PT-42	m	86	Prostate	1x8 Gy
PT-43	f	68	Breast	1x8 Gy
PT-44	m	57	Lung	1x8 Gy
PT-45	f	78	Lung	5x4 Gy
PT-46	m	82	Prostate	1x8 Gy
PT-47	m	64	Esophagus	1x8 Gy
PT-48	f	68	Breast	1x8 Gy
PT-49	f	75	Breast	1x8 Gy
PT-50	m	66	Prostate	1x8 Gy
PT-51	f	69	Breast	1x8 Gy
PT-52	m	57	Multiple Myeloma	5x4 Gy
PT-53	m	77	Lung	1x8 Gy
PT-54	f	58	Breast	1x8 Gy



Chapter 3

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

Published in

Bone

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 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.

Keywords: Bone fracture risk, femur, finite element, deep learning, lesion segmentation

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 kV, 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).

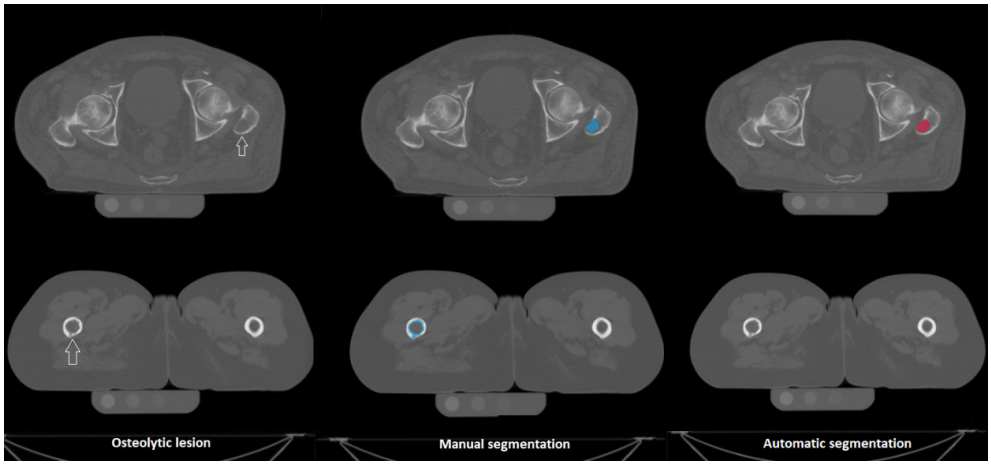


Fig 1: Osteolytic lesion and segmentation visualization.
Upper row shows an osteolytic lesion in the greater trochanter (white arrow) and the manual segmentation (mask in blue) which matches with the automatic segmentation (mask in red). This example only compared the manual and automatic segmentation 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).

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

FE-no-SL group failure loads ranged from 2839N to 10965N (mean 6295N), whereas failure loads obtained from FE-with-SL group ranged from 2124N to 10965N (mean 6047N). 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).

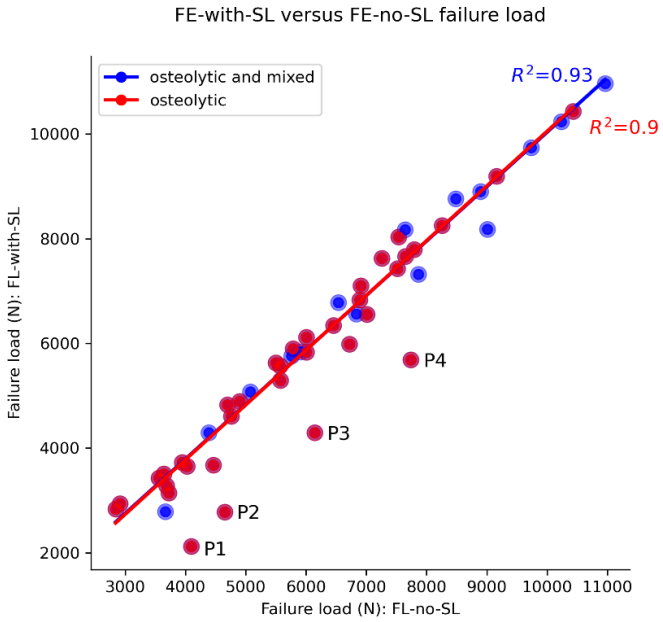


Fig 2: The correlation between failure loads obtained from femurs in the FE-with-SL group, and femurs in the FE-no-SL group for osteolytic and mixed lesions (n=50; red and blue dots) and osteolytic lesions (n=35; red dots). P1-4 are patients clearly deviating from the regression lines.

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 (diff_{P1}=48%, diff_{P2}=40%, diff_{P3}=30%, and 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 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.

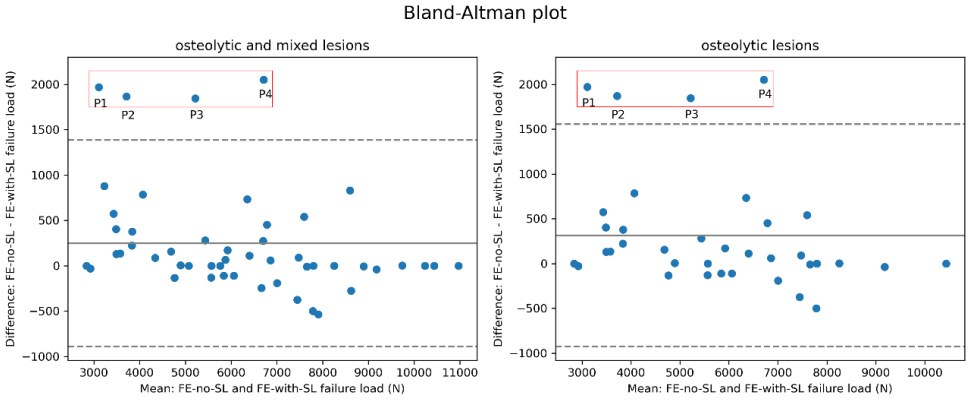


Fig 3: Bland-Altman plot representing the failure load difference between the FE-no-SL and FE-with-SL groups against the mean failure load for patients with osteolytic and mixed lesions (left), and patients with osteolytic lesions (right). Red boxes include femurs with the most deviating failure load difference from the average line (P1-4).

Lesion size ranged from 0 to 44321 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.

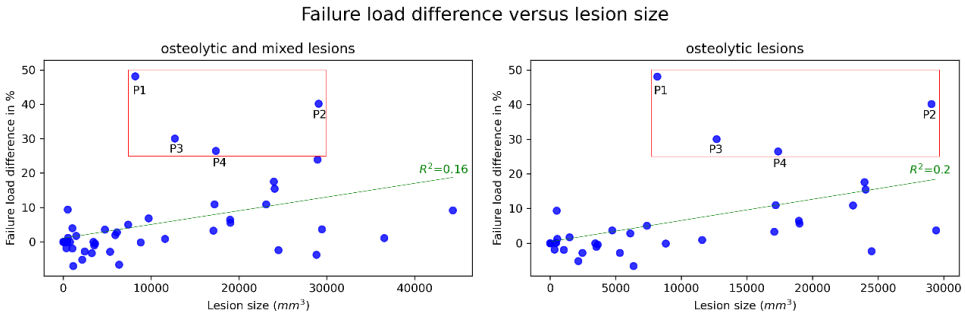


Fig 4: The correlation between failure load difference (100% * (FL-no-SL minus FE-with-SL)/FE-no-SL) and lesions size of the osteolytic and mixed group (left), and osteolytic group (right). Red boxes include femurs with the most deviating failure load difference from the regression line (P1-4). The green lines indicate linear regression.

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.

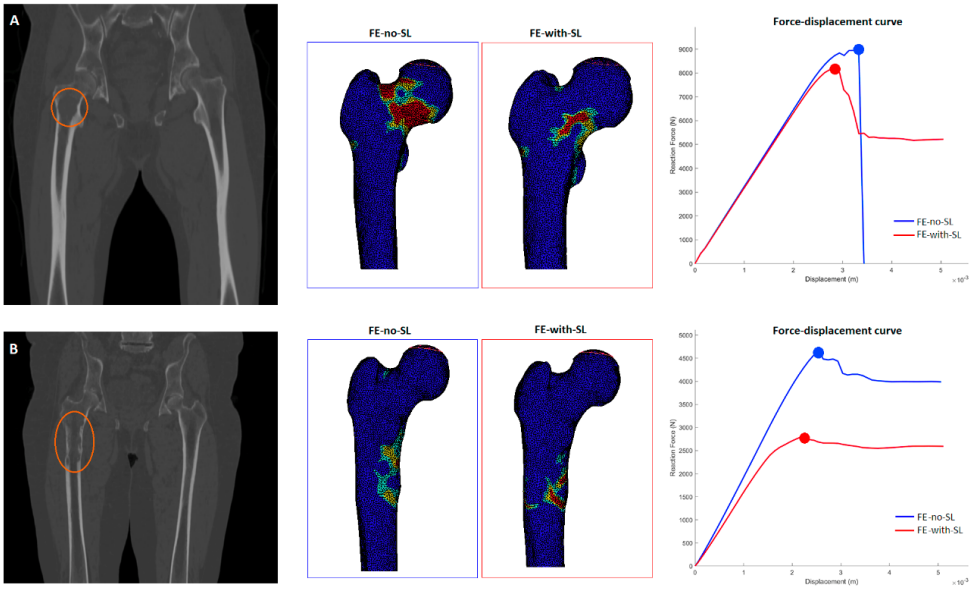


Fig 5: Femoral osteolytic lesions of two patients are shown (coronal view - orange circle).
Failure patterns (plasticity status) obtained from FE-no-SL model (blue box) and FE-with-SL model (red box) are visualized. The plasticity status is shown at the maximal load (red and blue dot) in the Force-displacement curve (right hand side).

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.

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%

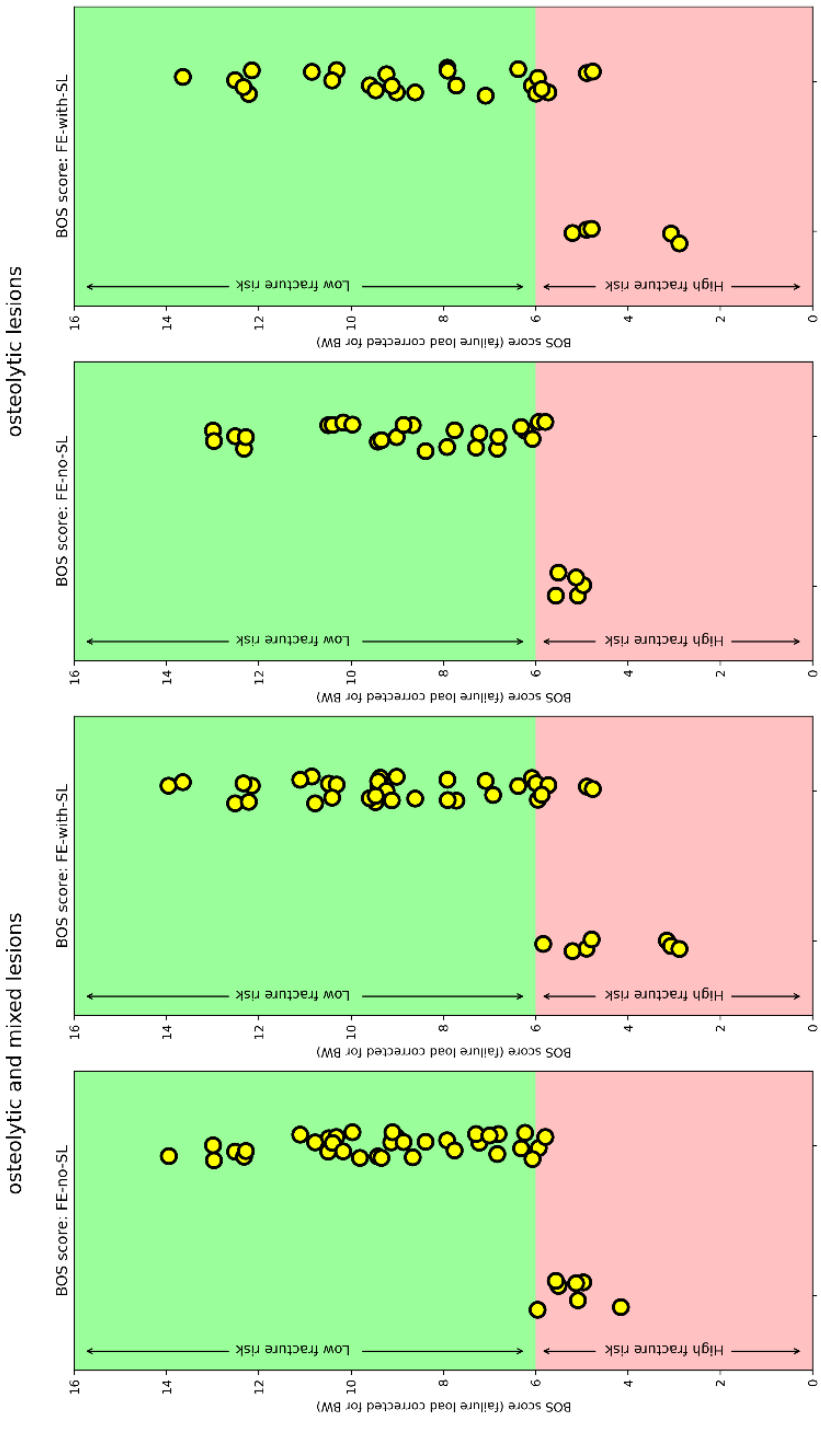


Fig 6: Each yellow dot represents a femur in the database. The graph provides comparison between FE-no-SL and FE-with-SL BOS scores (failure loads corrected for patients body weight [9]) of the patients with osteolytic and mixed lesions, and osteolytic lesions, respectively. Each graph is divided to 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.

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, postprocessing, 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 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.

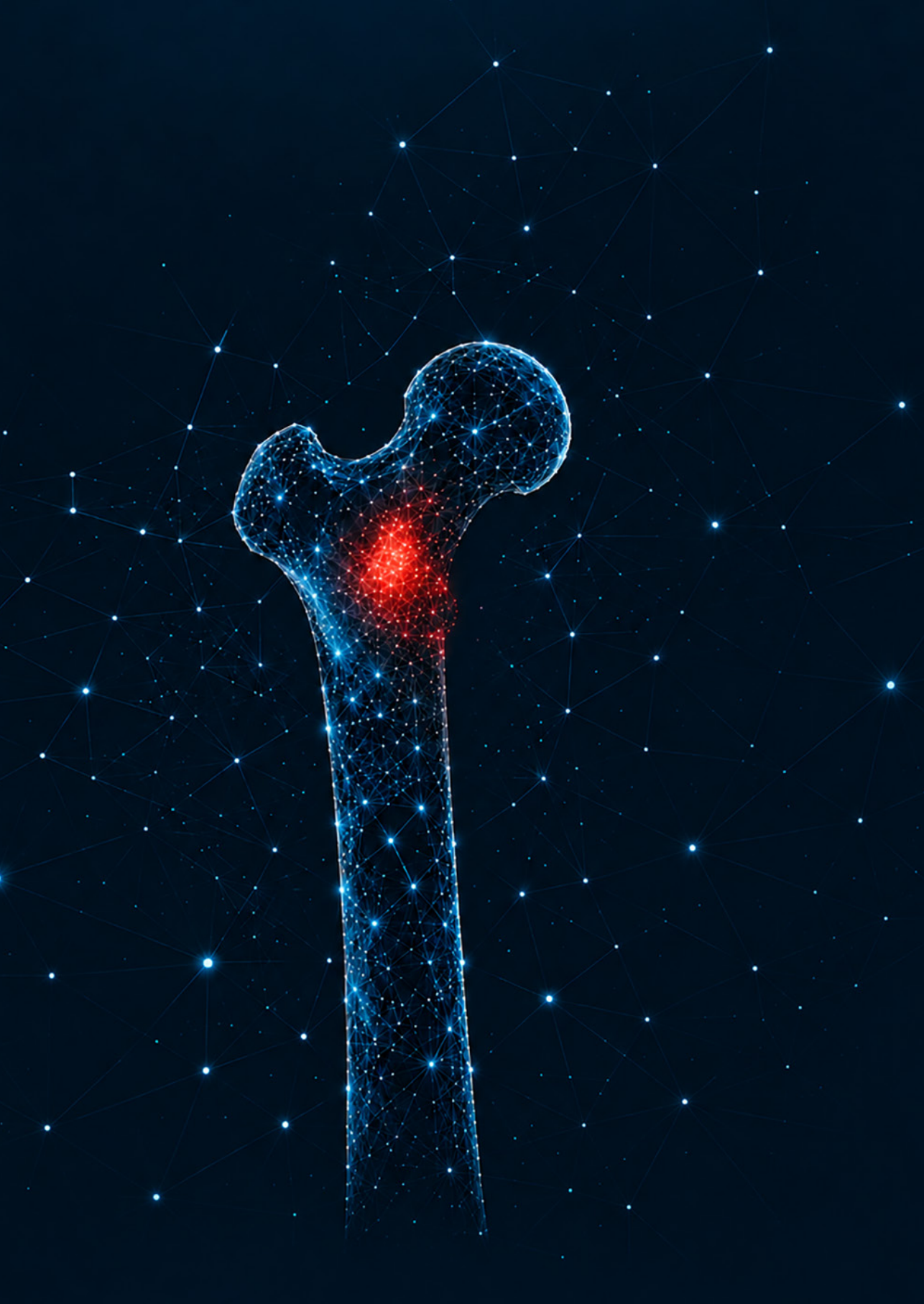
5. CONCLUSION

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.

6. REFERENCES

1. Macedo, F., et al., *Bone Metastases: An Overview.* Oncol Rev, 2017. **11**(1): p. 321.
2. Gralow, J.R., et al., *NCCN task force report: bone health in cancer care.* Journal of the National Comprehensive Cancer Network, 2009. **7**(Suppl_3): p. S-1-S-32.
3. Ataei, A., et al., *Evaluation of inter- and intra-operator reliability of manual segmentation of femoral metastatic lesions.* International Journal of Computer Assisted Radiology and Surgery, 2021.
4. Chambers, A.F., et al., *Critical Steps in Hematogenous Metastasis: An Overview.* Surgical Oncology Clinics of North America, 2001. **10**(2): p. 243-255.
5. Axelrod, D., A.M. Gazendam, and M. Ghert, *The Surgical Management of Proximal Femoral Metastases: A Narrative Review.* Current Oncology, 2021. **28**(5): p. 3748-3757.
6. Angelini, A., et al., *Treatment of pathologic fractures of the proximal femur.* Injury, 2018. **49**: p. S77-S83.
7. Eggermont, F., et al., *Can patient-specific finite element models better predict fractures in metastatic bone disease than experienced clinicians?* Bone & Joint Research, 2018. **7**(6): p. 430-439.
8. Eggermont, F., et al., *Patient-specific finite element computer models improve fracture risk assessments in cancer patients with femoral bone metastases compared to clinical guidelines.* Bone, 2020. **130**: p. 115101.
9. Eggermont, F., et al., *A Patient-Specific Fracture Risk Assessment Tool for Femoral Bone Metastases: Using the Bone Strength (BOS) Score in Clinical Practice.* Cancers, 2022. **14**(23): p. 5904.
10. Stadelmann, M.A., et al., *Conventional finite element models estimate the strength of metastatic human vertebrae despite alterations of the bone's tissue and structure.* Bone, 2020. **141**: p. 115598.
11. Kaneko, T.S., et al., *Mechanical properties, density and quantitative CT scan data of trabecular bone with and without metastases.* Journal of biomechanics, 2004. **37**(4): p. 523-530.
12. Milletari, F., N. Navab, and S.-A. Ahmadi. *V-net: Fully convolutional neural networks for volumetric medical image segmentation*. in *2016 Fourth International Conference on 3D Vision (3DV)*. 2016. IEEE.
13. Kooi, T., et al., *Large scale deep learning for computer aided detection of mammographic lesions.* Med Image Anal, 2017. **35**: p. 303-312.
14. Wang, J., et al., *A multi-resolution approach for spinal metastasis detection using deep Siamese neural networks.* Computers in biology and medicine, 2017. **84**: p. 137-146.
15. Chmelik, J., et al., *Deep convolutional neural network-based segmentation and classification of difficult to define metastatic spinal lesions in 3D CT data.* Med Image Anal, 2018. **49**: p. 76-88.
16. Roth, H.R., 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*. 2015, Springer. p. 3-12.
17. Madabhushi, A. and G. Lee, *Image analysis and machine learning in digital pathology: Challenges and opportunities.* Med Image Anal, 2016. **33**: p. 170-175.
18. Isensee, F., et al., *nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation.* Nature methods, 2021. **18**(2): p. 203-211.
19. Rolnick, D., et al., *Deep learning is robust to massive label noise. arXiv 2017.* arXiv preprint arXiv:1705.10694.

20. Eggermont, F., 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, 2019. **14**(7): p. e0220564.
21. Ronneberger, O., P. Fischer, and T. Brox. *U-net: Convolutional networks for biomedical image segmentation*. in *International Conference on Medical image computing and computer-assisted intervention*. 2015. Springer.
22. Refaeilzadeh, P., L. Tang, and H. Liu, *Cross-validation.* Encyclopedia of database systems, 2009. **5**: p. 532-538.
23. Keyak, J.H., et al., *Predicting Proximal Femoral Strength Using Structural Engineering Models.* Clinical Orthopaedics and Related Research®, 2005. **437**: p. 219-228.
24. Fleps, I. and E.F. Morgan, *A Review of CT-Based Fracture Risk Assessment with Finite Element Modeling and Machine Learning.* Current Osteoporosis Reports, 2022. **20**(5): p. 309-319.
25. Paranavithana, I.R., et al., *Systematic Review of Tumor Segmentation Strategies for Bone Metastases.* Cancers, 2023. **15**(6): p. 1750.
26. Fang, Y., et al., *The impact of training sample size on deep learning-based organ auto-segmentation for head-and-neck patients.* Physics in Medicine & Biology, 2021. **66**(18): p. 185012.
27. Orlando, N., et al., *Effect of dataset size, image quality, and image type on deep learning-based automatic prostate segmentation in 3D ultrasound.* Physics in Medicine & Biology, 2022. **67**(7): p. 074002.
28. Leclerc, S., 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)*. 2018. IEEE.
29. Narayana, P.A., et al., *Deep-learning-based neural tissue segmentation of MRI in multiple sclerosis: effect of training set size.* Journal of Magnetic Resonance Imaging, 2020. **51**(5): p. 1487-1496.
30. Ning, H., et al., *Choosing an appropriate training set size when using existing data to train neural networks for land cover segmentation.* Annals of GIS, 2020. **26**(4): p. 329-342.
31. Ciga, O. and A.L. Martel, *Learning to segment images with classification labels.* Medical Image Analysis, 2021. **68**: p. 101912.
32. Rajapakse, C.S., et al., *Influence of bone lesion location on femoral bone strength assessed by MRI-based finite-element modeling.* Bone, 2019. **122**: p. 209-217.
33. Keyak, J.H., et al., *Predicting the strength of femoral shafts with and without metastatic lesions.* Clinical Orthopaedics and Related Research (1976-2007), 2005. **439**: p. 161-170.
34. Damron, T.A. and K.A. Mann, *Fracture risk assessment and clinical decision making for patients with metastatic bone disease.* Journal of Orthopaedic Research®, 2020. **38**(6): p. 1175-1190.
35. Derikx, L.C., 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.* The Journal of bone and joint surgery. British volume, 2012. **94**(8): p. 1135-1142.
36. Whyne, C.M., S.S. Hu, and J.C. Lotz, *Burst fracture in the metastatically involved spine: development, validation, and parametric analysis of a three-dimensional poroelastic finite-element model.* Spine, 2003. **28**(7): p. 652-660.
37. Costa, M., et al., *Biomechanical assessment of vertebrae with lytic metastases with subject-specific finite element models.* Journal of the mechanical behavior of biomedical materials, 2019. **98**: p. 268-290.



Chapter 4

Assessing the impact of density-adapted osteoblastic lesion in metastatic femur using finite element analysis and deep learning-based lesion segmentation

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

ABSTRACT

Finite element (FE) models which are automatically generated from CT scans can be used to predict fracture risk in metastatic femora by calculating a BOne Strength (BOS) score. However, with osteoblastic lesions, reliable prediction is challenging due to the high bone density in the lesions, thereby overestimating bone strength. Hence, osteoblastic properties should be adapted and for this, lesions need to be separately identified. However, manual segmentation of these lesions is a time-consuming and subjective task, while automatic segmentation using deep learning algorithms requires validation. This study aimed to assess how modifying the bone mineral density (BMD) values of osteoblastic lesions affects the (BOS) score, evaluate the reliability of automatic segmentation, and compare BOS scores derived from manual and automatic segmentation.

Nineteen femora with osteoblastic metastases were analyzed using patient-specific FE models. The BMD of the metastases was adjusted in steps, reducing strength incrementally from 100% to 0%. The BOS score was calculated for each step to assess its sensitivity to BMD distribution changes. Manual and automatic segmentations were performed, and their agreement was evaluated using Dice coefficients. Additionally, Bland-Altman and correlation analyses were conducted to compare BOS scores resulting from the two segmentation methods.

The BOS score generally decreased as the BMD of osteoblastic metastases was reduced. However, the relationship was not always gradual, with some femora showing minimal changes when BMD was decreased. Dice coefficients for automatic segmentation varied widely (0 to 0.88), with lower accuracy in lesions located in the femoral trochanter and diaphysis level. BOS scores based on manual and automatic segmentation were strongly correlated, and in most cases the difference were relatively small. Six femora had higher BOS scores for manual segmentation, eight for automatic, and five showed virtually no difference. In most cases, the difference were small and not likely to be clinically relevant.

In conclusion, adjusting the BMD distribution of osteoblastic lesions influences BOS scores, but the effect varies based on lesion characteristics such as size and location. Automatic segmentation showed promise but exhibited variability depending on lesion location. Future research should focus on refining BMD adaptation for improved fracture risk prediction and expanding the training dataset for automatic segmentation to enhance its reliability.

INTRODUCTION

The skeleton is the third most common site for cancer metastases [1]. There are three subcategories of bone metastases, classified based on their radiographic presentation and their interaction with normal bone remodeling processes: osteoblastic, osteolytic, and mixed metastases [2]. Osteoblastic metastases are characterized by the formation and deposition of new bone [2-4]. Conversely, osteolytic metastases are associated with bone resorption and destruction [2-4]. Mixed metastases contain elements of both osteoblastic and osteolytic types [2, 4].

Pathological fractures represent a significant skeletal complication for patients with bone metastases [2]. The most vulnerable locations for these fractures are the spine and the long bones, with the femur being the most frequently affected site [4]. Given the profound impact on mobility and quality of life, accurate prediction of the risk of femoral fractures is crucial in determining the most effective treatment strategy [2]. One common approach is the Mirels scoring system, which evaluates four factors: location, type and size of metastases, and patient-reported pain, using a 12-point scale [5-7]. Mirels recommended surgical stabilization for patients scoring nine or above [6, 7]. However, studies by Van der Linden et al. and Van der Wal et al. have shown that reliance on the Mirels scoring system can lead to over-treatment of patients [5, 6]. Another predictor of fracture risk is the extent of axial cortical involvement of the bone metastases, determined on conventional radiographs [5, 6, 8]. Yet, this approach also has limitations, with a positive predictive value (PPV) of 20-23%, indicating a high potential for over-treatment [5, 6].

Patient-specific finite element (FE) computer modeling offers an alternative and promising approach for predicting fracture risk in femora affected by bone metastases. This method takes into account a range of factors including mechanical strength, bone mineral density (BMD) distribution, as well as the location and geometry of both the bone metastases and the femur itself [9]. FE models are generated from computed tomography (CT) scans, which provide detailed information on bone geometry and BMD distribution. From this data, mechanical properties are deduced. The fracture risk is then assessed by virtually loading the FE model until the femur fractures, at which point the corresponding failure load is determined [9]. This failure load is divided by the patient's body weight to derive the BOne Strength (BOS) score, a metric used for fracture risk prediction [10]. The BOS score is categorized into risk levels based on set thresholds: scores above eight indicate low fracture risk, scores below 6.5 suggest high risk, and scores between 6.5 and 8 are indicative of moderate risk.

Eggermont et al. compared the diagnostic accuracy of their FE modeling framework with axial cortical involvement in predicting fracture risk in femora with metastases [11]. They found that the FE model improved fracture risk prediction accuracy compared to axial cortical

involvement, reducing unnecessary surgeries from 81% to 61% [11]. Additionally, in another study, they demonstrated the clinical impact of their newly developed BOS score [10]. The BOS score significantly aided physicians and patients in making shared treatment decisions for femoral metastases [10].

Currently, the BOS score is applicable only to osteolytic metastases. A significant challenge arises in the use of FE modeling for predicting fracture risk in femora with osteoblastic metastases. These metastases have an increased density on CT scans, but are structurally weak, leading to lower strength than the BMD would suggest [12, 13]. This results in an overestimation of the strength of osteoblastic metastases by the FE model and a biased BOS score [10, 11, 14].

To address this issue of distorted mechanical properties in FE models for osteoblastic metastases, a viable approach is to adapt the BMD of the metastases. This modification aims to align the mechanical properties of the metastases more closely with reality, ultimately correcting the overestimated strength [10, 14, 15].

Adjusting the BMD in specific FE elements necessitates precise segmentation of osteoblastic metastases from femoral CT scans. The traditional approach is manual segmentation, conducted by skilled personnel. We earlier explored the consistency and reliability of this method for osteolytic, osteoblastic, and mixed lesions, both between different operators and within the same operator's work [15]. The findings revealed that manual segmentation presents significant challenges, yielding suboptimal reliability scores in both inter- and intra-operator contexts [15]. Moreover, this manual process of metastatic lesions segmentation is not only time-consuming but also demands substantial labor.

An alternative solution may lie in the use of automated segmentation, for instance using a deep learning algorithm. This type of artificial intelligence learns patterns from training data, enabling it to apply these insights to specific tasks. Within the scope of metastasis segmentation, a deep learning algorithm would be trained using data from manual segmentations to learn the segmentation process. It is crucial that such automated segmentations are both reliable and comparable with manual segmentations in terms of accuracy.

With this study we intend to enhance the applicability of the BOS score to patients with osteoblastic lesions and aimed to:

1) Assess how adapting material properties into manually segmented osteoblastic lesions affects the failure load and the resulting BOS score.
2) Explore the feasibility of employing an automatic segmentation method for osteoblastic lesions and evaluate its reliability.
3) Compare the differences in failure load and BOS score between manually and automatically segmented osteoblastic lesions when applying realistic material properties to them.

METHODS

Patient population

Two multicenter prospective cohort studies were conducted, one between 2006 and 2009 and the other between 2015 and 2017 [11, 14]. The purpose of these studies was to investigate the efficacy of FE modeling in predicting fracture risk in patients with femoral metastases. Patients undergoing palliative radiotherapy for femoral metastases at one of four Dutch radiotherapy institutes (Radboud university medical center, Nijmegen; Radiotherapeutic Institute Friesland, Leeuwarden; Bernard Verbeeten Institute, Tilburg; Leiden University Medical Center, Leiden) were invited to participate in the studies. Ethical approval was secured from all four institutes, and informed consent was obtained from each participating patient. To allow radiotherapy planning, QCT scans of the femoral area were obtained from the patients. These QCT scans were also utilized for generating the FE models [11, 14].

For the current study, sixteen patients with predominantly femoral osteoblastic metastases were selected from the cohort of the two multicenter prospective studies. Among these, thirteen patients had osteoblastic metastases in one femur, and three had osteoblastic metastases in both femora, totaling nineteen femora for the current dataset. The baseline characteristics of the patients can be found in Table 1.

Table 1. Patient characteristics.
KPS: Karnofsky Performance Status (0 = lowest performance (deceased), 100 = highest performance (no complaints, no evidence of disease)); Pain Score: 0 = no pain, 10 = worst pain imaginable; *RT*: Radiotherapy Schedule.

Patient number	Side	Weight (kg)	Height (cm)	Age	Sex	KPS	Pain score	Primary tumor	RT Schedule	Time till death (months)
1	L	68	180	66	M	60	8	Prostate	1x8 Gy	5
2	L	80	178	55	M	70	0	Prostate	1x8 Gy	unknown
3	R	68	181	70	M	70	6	Prostate	1x8 Gy	9
4	R	65	182	72	M	90	2	Prostate	1x8 Gy	23
5	R	80	181	75	M	70	8	Prostate	5x4 Gy	2
6	R	74	187	48	M	60	3	Lung	1x8 Gy	4
7	L+R	85	175	79	M	70	3	Prostate	1x8 Gy	6
8	L	82	174	83	M	80	7	Prostate	5x4 Gy	7
9	R	92	171	58	F	80	6.5	Breast	1x8 Gy	unknown
10	L	118	190	65	M	80	7	Prostate	unknown	30
11	R	63	151	59	F	70	5	Breast	1x8 Gy	0
12	R	101	181	66	M	80	4	Prostate	1x8 Gy	8
13	L	88	186	75	M	70	6	Prostate	1x8 Gy	19
14	R	124	unknown	55	M	80	8	Prostate	1x8 Gy	11
15	L+R	70	175	65	M	70	5	Prostate	1x8 Gy	13
16	L+R	79	182	95	M	80	3	Prostate	No RT	unknown

Lesion segmentation

The osteoblastic metastases were segmented, using both manual and automatic segmentation techniques.

For manual segmentation, an experienced operator delineated the osteoblastic metastases from the femoral CT scans using Mimics 14.0 (Materialise, Leuven, Belgium). This process involved the operator identifying the metastases and selecting their corresponding voxels (Figure 1).

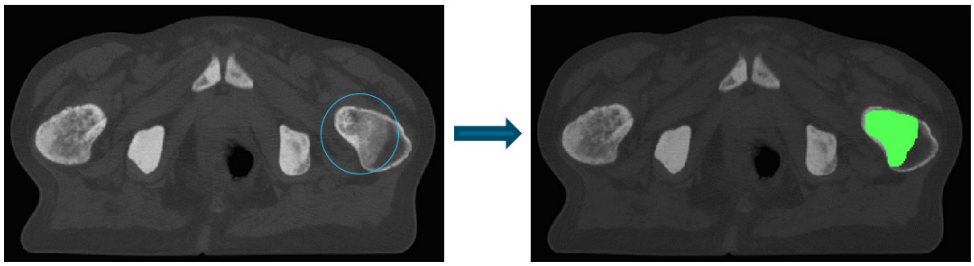


Figure 1. Illustration of manual segmentation of an osteoblastic lesion.

For automatic segmentation of the osteoblastic metastases a convolutional neural network (CNN) called nnU-net was employed, a deep learning tool specialized in analyzing visual imagery [16], which was previously applied to segment osteolytic bone metastases [17]. The nnU-net is notable for its self-configuring capability, which includes preprocessing datasets, setting parameters for the CNN, and applying post-processing [16]. The training dataset for the nnU-net comprised 119 CT scans [11, 14] of manually segmented femoral metastases, including osteoblastic (n = 50), osteolytic (n = 57), and mixed (n = 12) types. These scans were divided into left and right femur sub-volumes. Sub-volumes containing lesions and their corresponding manual segmentations were used to feed the automatic configuration and learning process of the nnU-net, leading to the development of trained models. These models were then applied to automatically segment osteoblastic metastases in the patient cohort of this study. The 16 patient selected for the material-adaptation analysis were part of the 119 CT scans used as the training dataset. For each patient, the network was trained on the CT scans of all other patients in the training dataset and then used to segment that patient, ensuring that no data from the test patient leaked into training.

FE models

FE models were generated for the nineteen femora (see Table I) with osteoblastic metastases based on a previously described workflow [9-11, 14] (Figure 2). The QCT scans provided the

three-dimensional geometry of the bones which was then transformed into a solid mesh made of four-node tetrahedral elements. To accurately determine the BMD of the femora, a solid calibration phantom with known calcium equivalent densities was scanned along with the femora during the QCT scans. This allowed for the conversion of Hounsfield Units (HU) from the QCT scans into BMD values.

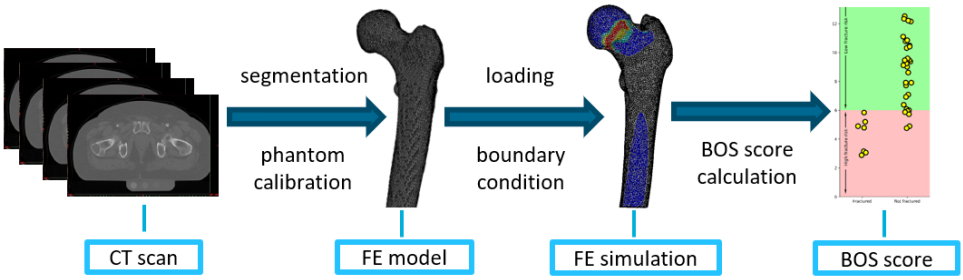


Figure 2. The FE simulation workflow starts by obtaining the model geometry from a CT scan, with bone mineral densities extracted through phantom calibration. Subsequently, boundary conditions, including fixation point at the knee joint center and loading, were applied. BOS score was calculated based on the patients body weight.

Based on these BMD values, the ash densities and non-linear isotropic material behavior was calculated for each tetrahedral element using the model as proposed by Keyak et al. [18]. Hence, this constitutive model describes the material behavior of each element, beginning with an elastic phase, followed by a perfectly plastic phase, followed by a strain softening phase, and culminating in an indefinite perfectly plastic phase [18].

The orientation of the bone models within 3D space were aligned to replicate the standing phase of walking, aligning the hip joint center (HJC) with the knee joint center (KJC). The models were constrained distally at the KJC using two high stiffness springs. During the simulation, the FE models were subjected to incremental axial loading using a cup placed on the femoral head. The output of the simulation was the failure load at which the femur virtually fractured (i.e. the maximal force) measured in Newtons (N) [9, 11, 14]. The fracture location was identified by identifying the elements in the plastic state at the at failure load [9]. The BOS score was then calculated by dividing the failure load by the patient's body weight. The simulations were conducted using Marc Mentat 2021.4 (Hexagon AB, Stockholm, Sweden).

Outcome measures

To address the first objective of this study, we investigated how modifying the material properties of osteoblastic metastases affected the BOS score. In our study, we did not directly reduce the yield strength of the osteoblastic lesions but instead reduced the BMD values of the lesion voxels. Based on the material model by Keyak et al. [18], reduction in BMD leads to lower ash density, which also results in a lower yield strength and lower Young's modulus. Therefore, the BMD reduction indirectly accounts for decreased mechanical properties of the osteoblastic lesions. We created a series of FE models where the BMD values of these lesion voxels were reduced incrementally, corresponding to reduction in yield strength of 0%, 20%, 40%, 60%, 80%, and 100% with respect to the original yield strength values. By comparing the BOS scores across these models, we assessed the sensitivity of different material property assumptions for the osteoblastic tissue. Additionally, the regions indicating plastic deformation were visualized to compare fracture locations.

For the second objective, we explored the feasibility and reliability of employing an automatic deep learning-based segmentation method for osteoblastic lesions. To evaluate the similarity between manual and automatic segmentation methods, we calculated the Dice coefficients of the obtained lesions for each femur, which quantifies the overlap between the manual and automatic segmentations. Additionally, we analyzed the potential relationship between segmentation volume and Dice coefficient to determine whether larger manual segmentation volumes corresponded to higher Dice coefficients, providing further insight into segmentation performance trends. This allowed us to assess how well the automatic method replicated the manual segmentation results.

For the third objective, we compared the BOS scores between the automatically and manually segmented osteoblastic lesions while implementing adapted mechanical properties to the osteoblastic lesion. However, there is limited data available on the strength of osteoblastic metastases, and the available data show variability in their mechanical properties. Stadelmann et al. [19] reported that osteoblastic metastases have higher bone density, but slightly lower tissue properties compared to non-involved surrounding bone, suggesting that their mechanical impact is not uniform. Based on the decrease in mechanical properties reported by Stadelmann et. al and supported by our own sensitivity analysis (as defined in the first research objective), we chose to reduce the BMD to a value that corresponded to 70% of its original Young's modulus (equal to 70.8% of the yield strength). The Bland-Altman method and correlation analysis were used to demonstrate the differences and correlations in BOS scores between automatic and manual segmentations.

RESULTS

Impact of material properties on BOS score

The effect of strength reduction of the osteoblastic lesions on the total femoral bone strength was highly variable within the group of 19 femora. When comparing the BOS scores between 0% and 100% strength reduction, 8 out of 19 femora showed changes within ±1 BOS score, indicating relatively minor variation. In contrast, the remaining 11 femora experienced a considerable decrease in BOS scores beyond -1. It was also seen that most femora exhibit decreasing trends in BOS scores as the strength of osteoblastic metastases decreases (Figure 3), although some femora remained their strength even when the osteoblastic lesion was modeled without any strength.

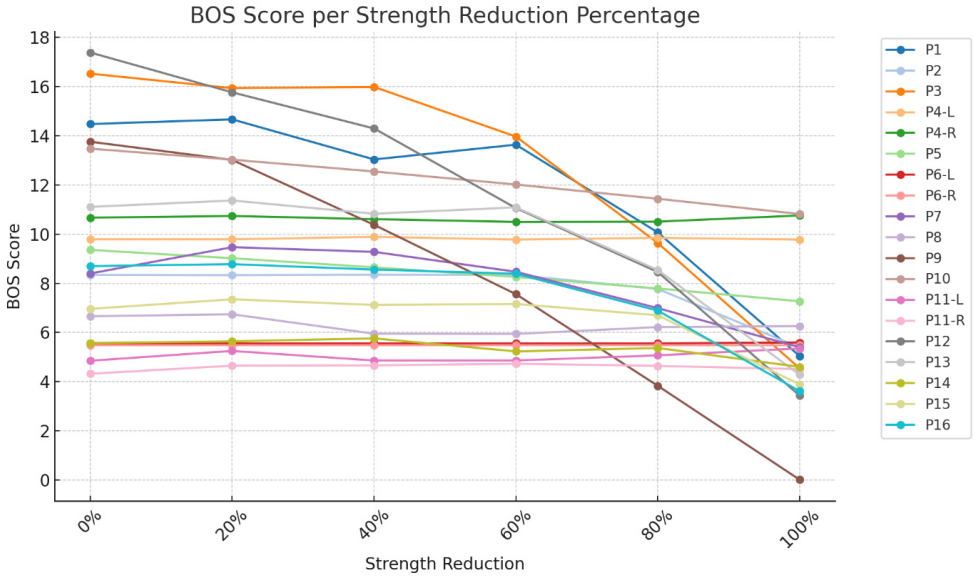


Figure 3. BOS scores (y-axis) for all femora across various strength reduction percentages (x-axis). Each line represents an individual femur, with dots marking the corresponding strength calculations.

The reduction in the strength percentage of the lesions led to a variety of responses, showing only slight to substantial changes in the plasticity region, with some cases showing a complete shift in the plasticity area. For clarity, two examples were selected to illustrate how variations in the lesion size and location affected the BOS score and plasticity pattern across different stages of lesion strength reduction (Figure 4). The change in the plasticity region was influenced by both the size and location of the lesions. In the first case (P6-L), where the lesions were small and more spread along the femoral shaft, the plasticity region remained stable, and the BOS score

showed minimal change. In contrast, the second case (P7), with a larger lesion near the femoral trochanter, showed a more noticeable shift in plasticity and a greater BOS score reduction as the lesion strength decreased.

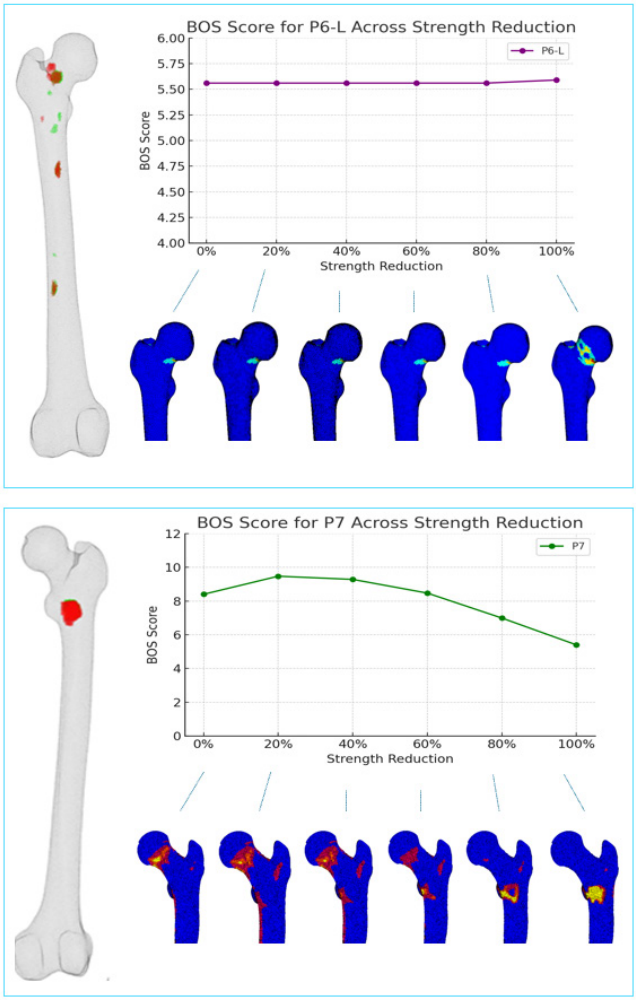


Figure 4. Shows two distinct cases of BOS score response to strength reduction. For P6-L (top), the BOS score remained consistent across different strength levels, with plasticity regions unaffected due to smaller, evenly distributed metastases along the femur. Conversely, P7 (bottom) demonstrated a BOS score decline and shifting plasticity regions, influenced by larger metastases. Red and green areas within the left larger femora indicate regions where the osteoblastic lesions are segmented manually and automatically, respectively.

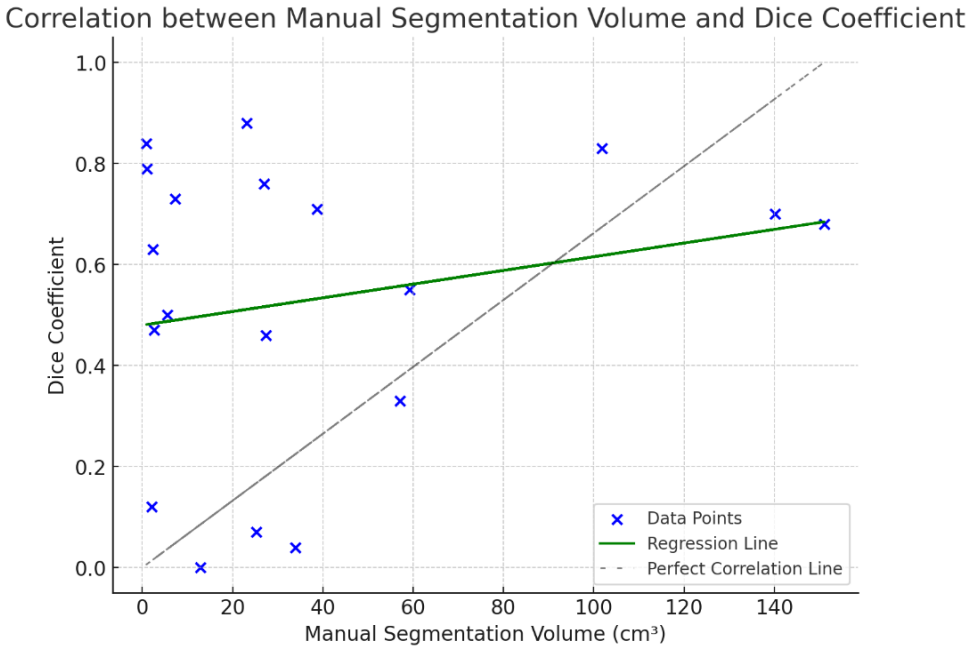


Figure 5. The correlation between manual segmentation volume and Dice coefficient. A weak overall correlation was seen (green trend line), implying that increase in lesion volumes do not necessarily correspond to increase in Dice coefficients.

Comparison of BOS score: manual versus automatic segmentation

The third part of the study aimed to assess the differences and relationships between BOS scores obtained through manual and automatic segmentation methods. The majority of the femora showed marginal differences between -1 and 0.5 in BOS score between automatic and manual segmentation methods. In two out of the nineteen femora, the BOS score obtained with manual segmentation was much smaller than the BOS score based on automatic segmentation. Six femora showed a higher BOS score for manual segmentation, eight had higher scores for automatic segmentation, and five showed no difference between the two segmentation methods (Figure 6). The correlation analysis (Figure 7) further demonstrated a strong relationship between manual and automatic BOS scores.

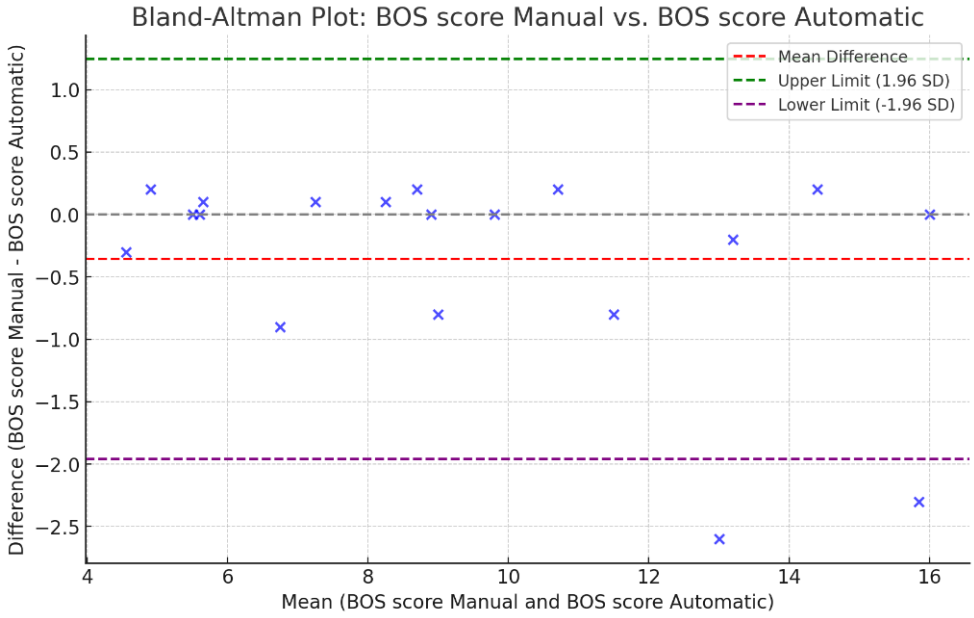


Figure 6. Showed Bland-Altman plot comparing BOS scores obtained from manual and automatic segmentation methods. Most points lie within the limits of agreement (green and purple dashed lines), with a distribution of differences: Six femora showed higher manual BOS scores, eight showed higher automatic BOS scores, and five showed no difference.

Feasibility and reliability of automatic segmentation

Across all the 19 femora, the mean Dice coefficient comparing each manual and automatic segmentation was 0.53 (SD=0.27), with values ranging from 0 to 0.88. The lesions within the femoral head region exhibited relatively high and consistent Dice coefficients, indicating reliable automatic lesion segmentation. In contrast, lesions located in the trochanter and diaphysis region showed a wider range of Dice coefficients with an outlier at the value equal to zero, suggesting variable segmentation quality. These findings highlight that segmentation involving the diaphysis is particularly challenging and prone to low accuracy, likely contributing to the observed variability in Dice coefficients.

Additionally, it was investigated whether larger manual segmentation volumes (being easier detectable) corresponded to higher Dice coefficients. A weak positive correlation between segmentation volume and Dice coefficient, but with considerable variability, was seen (Figure 5).

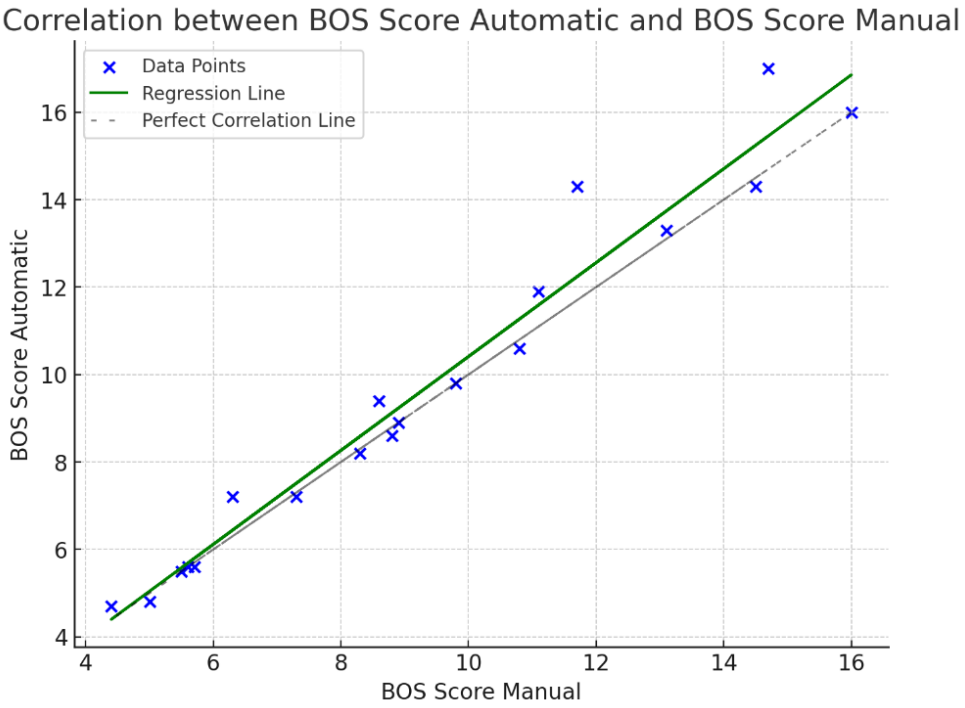


Figure 7. Correlation graph comparing BOS scores obtained from manual and automatic segmentation methods. Most points align closely with gray line (perfect correlation), indicating high agreement between the two methods.

The two most deviating points in Figure 6 and 7 reveal cases where segmentation differences and lesion location affect the BOS score. In the first case, a lesion at the femoral trochanter showed a difference in lesion volume of 56.5 cm^3 between manual and automatic segmentations (Dice 0.7). In the second case, lesions involving the femoral head, trochanter, and diaphysis had a difference in volume of 63 cm^3 between manual and automatic segmentations (Dice 0.68). Despite these relatively high differences in volume, the Dice scores were moderate, suggesting that critical parts of the lesions may not have been segmented in one method, leading to relatively high BOS score differences.

DISCUSSION

The first objective of this study was to assess how adapting material properties into manually segmented osteoblastic lesions impacts the resulting BOS score. The second aim was to explore the feasibility of employing an automatic segmentation method for osteoblastic lesions and

evaluate its reliability. Lastly, the study aimed to compare the differences in BOS score between manually and automatically segmented osteoblastic lesions when realistic material properties were applied. To achieve these objectives, nineteen femoral FE models with osteoblastic metastases were analyzed, with their BMD distributions modified using both manual and automatic segmentation methods.

The analysis addressing the first aim revealed that in the majority of femora, the BOS score decreased as the strength percentage of the osteoblastic metastases decreased. However, the decrease in BOS score was not consistently gradual with each reduction increment in strength. Remarkably, in the cases of eight femora (P4-L, P4-R, P6-L, P6-R, P8, P11-L, P11-R, and P14), the different strength percentages had relatively low (±1 BOS score) impact on the BOS score. This could potentially be attributed to variations in the size and location of the metastases. Obviously, if metastatic lesions are located at critical fracture-prone areas, the reduction in strength will have a larger effect than when the lesions are situated at a location where the bone is loaded marginally. The FE method used in this study is able to incorporate biomechanical aspects that affect the load-bearing capacity of metastatic femur. In contrast, commonly used clinical tools such as a scoring system suggested by Mirels [7] and assessment based on axial cortical involvement [5, 6, 8] do not fully capture these biomechanical factors. As a result, they may have limited accuracy in predicting actual fracture risk and often lead to over-treatment.

Interestingly, four femora (P4-R, P6-L, P11-L, and P11-R) exhibited a slight and unexpected increase in the BOS score which was calculated with the lowest bone strength in the lesions. One possible explanation is that the fracture pattern in these femora may have been altered by the presence of metastases, leading to fractures occurring in stronger, less vulnerable regions of the bone. This suggests that as local strength decreases, an alternative fracture pattern may emerge, where stronger elements converge into a fractured (plastic) state, therefore increasing resistance to femoral fracture. These observations highlight the complex interaction between metastases distribution and fracture mechanics in determining BOS scores.

Although some femora showed large BOS score differences between manual and automatic segmentation, the two methods were still strongly correlated. None of the femora had identical segmentation or volume measurements for the two methods. These differences in BOS scores can be explained by discrepancies in both segmentation volume and location. The variation is likely not caused by the neural network itself, but rather by the complexity of the lesion segmentation task. Osteoblastic lesions often have unclear borders, making it difficult even for experienced observers to agree on where the lesion begins and ends. Our previous inter- and intra-operator study also showed this variation between manual segmentations [15]. When the ground truth segmentations are uncertain, the network learns and subsequently generates segmentations with similar uncertainty. In this study, we used nnU-Net, a well-established and

widely used network in medical image segmentation. Its architecture and training strategy are based on state-of-the-art methods and have been successfully applied in recent studies on bone and tumor detection [16]. The network was trained on 119 CT scans of manually segmented femoral metastases. Although this number is relatively small, earlier studies have shown that deep learning-based segmentation can still perform reliably with similar dataset sizes when trained with well-annotated data [20, 21]. Additionally, our study included a diverse set of femora with varying metastasis volumes and locations, which helps to evaluate the robustness of the method.

In one case, the automatic segmentation failed to detect any metastasis, while manual segmentation identified a relatively small lesion (2.1 cm^3) located near the femoral cortex. A possible reason for this erroneous finding could be the lesion's small volume and its close proximity to the cortex, which may present a challenge for automatic segmentation [17]. Despite such limitations, the study provided a comparative analysis of automatic and manual segmentation in terms of BOS score and showed how variations in segmentation affect strength prediction. It also investigated the effect of adapting the BMD distribution of femoral osteoblastic metastases on the BOS score, contributing to the understanding of fracture risk in patients with different lesion characteristics.

This study had a number of limitations. First, clinically none of the femora in the patient population fractured, making it difficult to compare the biomechanical assessments as performed in this study to clinical reality. In principle three femora in our database had BOS scores below 6.5, regardless of the assumed strength of the lesion tissue (see Figure 3), which classifies them as being at high fracture risk. However, we do not have information about how active these patients were and whether they fully loaded their bones during daily activities. This makes it difficult to compare our mechanical strength predictions directly to the clinical outcome. Hence, measuring patient activity levels, as done in other research fields [22], could help address this knowledge gap in future studies

Another limitation of this study was the dependence on empirical adjustments in BMD distribution to represent the strength percentages of osteoblastic metastases. This approach was necessary due to the lack of reliable data on osteoblastic metastasis strength in existing literature. Kaneko et al. reported an average ultimate strength of 7.41 MPa and a BMD value of 158 mg/cm^3 for osteoblastic metastases in the trabecular bone of the distal femur [23]. However, the average BMD distribution was notably higher in this study (554.3 mg/cm^3). This discrepancy could be attributed to Kaneko's small sample size of only two donors. Hence, the 7.41 MPa ultimate strength value was considered unreliable for adjusting the BMD distribution in this study. Instead, we chose to reduce the BMD of the osteoblastic lesions voxels to 70% of their original values. This was based on the observed decrease in mechanical properties reported by

Stadelmann et. Al [19], combined with our own sensitivity analysis. This approach allowed us to simulate reduced mechanical behavior without relying on a single and uncertain strength value.

Future research should focus more on the clinical relevance of these calculations. To our knowledge, no fractures have been clinically observed in femora with osteoblastic metastases within our patient cohort, making it difficult to directly validate the BOS score predictions in these cases. Additionally, optimizing BMD distribution adaptation for improved fracture risk prediction will benefit from including fractured femora in future studies. Obtaining more realistic mechanical properties from donor femora with osteoblastic metastases, and expanding the training dataset for automatic segmentation with more manually segmented CT scans, remain important steps to improve segmentation accuracy and model input quality.

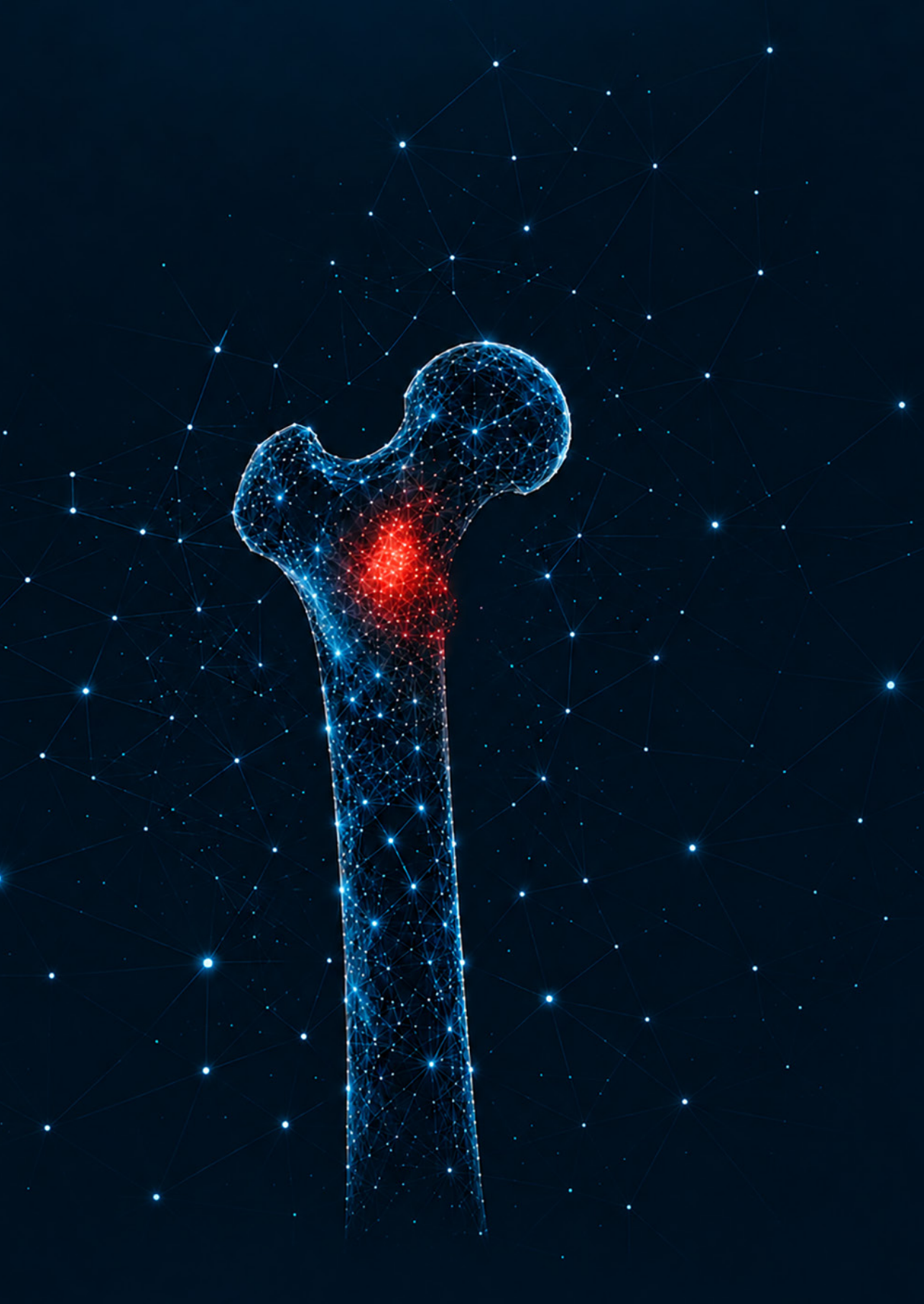
Finally, this study used one loading condition, but results showed that metastatic lesions along the typical fracture path can reduce bone strength. Other daily activities, such as stair climbing or standing up from a chair, involve different loading modes [24, 25], which may shift the critical stress regions. As a result, lesions that were not important in one loading scenario may become more critical in another. Future research should include multiple loading conditions to improve the generalizability of BOS score predictions.

In conclusion, adjusting the BMD distribution of osteoblastic lesions influences BOS scores, but the effect varies based on lesion characteristics. Automatic segmentation showed promise but exhibited variability depending on lesion location. Future research should focus on refining BMD adaptation for improved fracture risk prediction and expanding the training dataset for automatic segmentation to enhance its reliability.

REFERENCES

1. Johnson, S.K. and M.T. Knobf, *Surgical interventions for cancer patients with impending or actual pathologic fractures.* Orthopaedic Nursing, 2008. **27**(3): p. 160-171.
2. Coleman, R.E., *Skeletal complications of malignancy.* Cancer: Interdisciplinary International Journal of the American Cancer Society, 1997. **80**(S8): p. 1588-1594.
3. Mundy, G.R., *Metastasis to bone: causes, consequences and therapeutic opportunities.* Nature Reviews Cancer, 2002. **2**(8): p. 584-593.
4. Macedo, F., et al., *Bone metastases: an overview.* Oncology reviews, 2017. **11**(1).
5. Van der Wal, C., et al., *Axial cortical involvement of metastatic lesions to identify impending femoral fractures; a clinical validation study.* Radiotherapy and oncology, 2020. **144**: p. 59-64.
6. Van der Linden, Y., et al., *Comparative analysis of risk factors for pathological fracture with femoral metastases: results based on a randomised trial of radiotherapy.* The Journal of Bone & Joint Surgery British Volume, 2004. **86**(4): p. 566-573.
7. Mirels, H., *Metastatic Disease in Long Bones A Proposed Scoring System for Diagnosing Impending Pathologic Fractures.* Clinical Orthopaedics and Related Research (1976-2007), 1989. **249**: p. 256-264.
8. van der Linden, Y.M., et al., *Simple radiographic parameter predicts fracturing in metastatic femoral bone lesions: results from a randomised trial.* Radiotherapy and Oncology, 2003. **69**(1): p. 21-31.
9. Derikx, L.C., 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.* The Journal of Bone & Joint Surgery British Volume, 2012. **94**(8): p. 1135-1142.
10. Eggermont, F., et al., *A Patient-Specific Fracture Risk Assessment Tool for Femoral Bone Metastases: Using the Bone Strength (BOS) Score in Clinical Practice.* Cancers, 2022. **14**(23): p. 5904.
11. Eggermont, F., et al., *Patient-specific finite element computer models improve fracture risk assessments in cancer patients with femoral bone metastases compared to clinical guidelines.* Bone, 2020. **130**: p. 115101.
12. Roudier, M.P., et al., *Histopathological assessment of prostate cancer bone osteoblastic metastases.* The Journal of urology, 2008. **180**(3): p. 1154-1160.
13. Rikitake, H., et al., *Risk assessment of femoral pathological fracture in prostate cancer patients by computed tomography analysis.* Journal of Bone and Mineral Metabolism, 2022. **40**(4): p. 704-711.
14. Eggermont, F., et al., *Can patient-specific finite element models better predict fractures in metastatic bone disease than experienced clinicians?: Towards computational modelling in daily clinical practice.* Bone & joint research, 2018. **7**(6): p. 430-439.
15. Ataei, A., et al., *Evaluation of inter-and intra-operator reliability of manual segmentation of femoral metastatic lesions.* International journal of computer assisted radiology and surgery, 2021. **16**: p. 1841-1849.
16. Isensee, F., et al., *nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation.* Nature methods, 2021. **18**(2): p. 203-211.
17. Ataei, A., et al., *The effect of deep learning-based lesion segmentation on failure load calculations of metastatic femurs using finite element analysis.* Bone, 2024. **179**: p. 116987.

18. Keyak, J.H., et al., *Predicting proximal femoral strength using structural engineering models.* Clinical Orthopaedics and Related Research (1976-2007), 2005. **437**: p. 219-228.
19. Stadelmann, M.A., et al., *Conventional finite element models estimate the strength of metastatic human vertebrae despite alterations of the bone's tissue and structure.* Bone, 2020. **141**: p. 115598.
20. Narayana, P.A., et al., *Deep-learning-based neural tissue segmentation of MRI in multiple sclerosis: effect of training set size.* Journal of Magnetic Resonance Imaging, 2020. **51**(5): p. 1487-1496.
21. Ning, H., et al., *Choosing an appropriate training set size when using existing data to train neural networks for land cover segmentation.* Annals of GIS, 2020. **26**(4): p. 329-342.
22. Mooiweer, Y., M. Stevens, and I. Van den Akker-Scheek, *Being active with a total hip or knee prosthesis: a systematic review into physical activity and sports recommendations and interventions to improve physical activity behavior.* European Review of Aging and Physical Activity, 2022. **19**(1): p. 7.
23. Kaneko, T.S., et al., *Mechanical properties, density and quantitative CT scan data of trabecular bone with and without metastases.* Journal of biomechanics, 2004. **37**(4): p. 523-530.
24. Heller, M., et al., *Musculo-skeletal loading conditions at the hip during walking and stair climbing.* Journal of biomechanics, 2001. **34**(7): p. 883-893.
25. Taylor, W.R., et al., *Tibio-femoral loading during human gait and stair climbing.* Journal of orthopaedic research, 2004. **22**(3): p. 625-632.



Chapter 5

The effect of variations in CT scan protocol on femoral finite element failure load assessment using phantomless calibration

Ali Ataei [1], Jelle Eikhout [1], Ruud G. H. van Leeuwen [2], Esther Tanck [1], Florieke Eggermont [1]

[1]Orthopaedic Research Lab, Radboud Institute for Health Sciences, Radboud university medical center, P.O. Box 9101, 6500 HB Nijmegen, The Netherlands

[2]Department of Radiotherapy, Radboud Institute for Health Sciences, Radboud university medical center, Nijmegen, The Netherlands

Published in

Plos one

ABSTRACT

Recently, it was shown that fracture risk assessment in patients with femoral bone metastases using Finite Element (FE) modeling can be performed using a calibration phantom or air-fat-muscle calibration and that non-patient-specific calibration was less favorable. The purpose of this study was to investigate if phantomless calibration can be used instead of phantom calibration when different CT protocols are used. Differences in effect of CT protocols on Hounsfield units (HU), calculated bone mineral density (BMD) and FE failure loads between phantom and two methods of phantomless calibrations were studied.

Five human cadaver lower limbs were scanned atop a calibration phantom according to a standard scanning protocol and seven additional commonly deviating protocols including current-time, tube voltage, slice thickness, rotation time, field of view, reconstruction kernel, and reconstruction algorithm. The HUs of the scans were calibrated to BMD (in mg/cm^3) using the calibration phantom as well as using air-fat-muscle and non-patient-specific calibration, resulting in three models for each scan. FE models were created, and failure loads were calculated by simulating an axial load on the femur. HU, calculated BMD and failure load of all protocols were compared between the three calibration methods.

The different protocols showed little variation in HU, BMD and failure load. However, compared to phantom calibration, changing the tube voltage resulted in a relatively large decrease of approximately 10% in mean HU and BMD of the trabecular and cortical region of interest (ROI), resulting in a 13.8% and 13.4% lower failure load when air-fat-muscle and non-patient-specific calibrations were used, respectively.

In conclusion, while we observed significant correlations between air-fat-muscle calibration and phantom calibration as well as between non-patient-specific calibration and phantom calibration, our sample size was too small to prove that either of these calibration approaches was superior. Further studies are necessary to test whether air-fat-muscle or non-patient-specific calibration could replace phantom calibration in case of different scanning protocols.

Keywords: CT scan protocol, Calibration, Finite Element, Failure load

INTRODUCTION

Bone metastases occur in a large number of breast, prostate, thyroid, lung and kidney cancer patients [1]. These metastases can be painful and can increase the risk of pathological fractures.

Such fractures cause mobility problems resulting in the inability to perform activities of daily living, leading to a reduced quality of life and an increased mortality risk [2].

It is of great importance to accurately determine the fracture risk of these patients, so the appropriate treatment can be chosen. In current clinical practice, assessment of fracture risk is done using computed tomography (CT) scans or conventional radiography, however the assessment of the risk has been shown to be quite complex [3]. Mirels' scoring system is a commonly used method which calculates fracture risk based on the site and location of the lesion, as well as the appearance and severity of pain [4, 5]. However, several studies showed that the Mirels' score lacked specificity [3, 6-8]. Van der Linden et al. introduced a 30 mm axial cortical involvement as a simple criterion assisting clinicians to select the appropriate treatment [9], which was recently validated [10]. Using axial cortical involvement, specificity increased compared to Mirels' score whereas sensitivity was comparable [9, 10]. Since such assessments are mainly relying on the clinician's ability to interpret a radiographic image, patients might be either overtreated, involving unnecessary and cost-ineffective interventions, or might be undertreated, causing fractures and decreased quality of life because of further complications [3, 11].

Patient-specific finite element (FE) modeling can predict fracture risk with higher accuracy compared to the methods currently used by clinicians [11-14]. Input to the FE models can be CT scans [11, 13, 14] as well as MR images [15]. CT scans need to be calibrated to convert the Hounsfield units (HU) to bone mineral density (BMD) to calculate the mechanical properties of the bone. For this purpose, usually a calibration phantom [13, 14, 16-18] is scanned along with the patient. However, these phantoms are expensive and impractical. Furthermore, since they have to be scanned together with the patient, it is not possible to generate FE models of scans performed without a calibration phantom. In some studies, phantomless calibration methods are used to obtain patient-specific calibration functions, for example based on HU of intrinsic patient tissue such as fat, muscle, air or blood [19-24]. Such phantomless calibration methods appear to result in similar bone mineral densities compared to calibration with phantoms [16-22, 25]. However, there is limited information about how these methods affect failure load assessment using FE modeling. A few previous studies showed that their proposed methods of phantomless calibration, by using internal tissues to calculate a calibration function specific for each patient [20, 26] or by applying the same general calibration function for all patients, i.e. non-patient-specific calibration [27], was applicable to FE studies on retrospective cohorts lacking a calibration phantom [20, 26, 27].

Recently, Eggermont et al. developed a novel method that calibrates the CT scan based on the HU of air, fat, and muscle tissue and studied the influence of this air-fat-muscle calibration on femoral FE failure load assessment [19]. They found a very high correlation (R^2=0.94) between phantom calibration and air-fat-muscle method, and accordingly, no significant difference in

FE failure loads between the phantom calibration and the air-fat-muscle calibration. They also applied a non-patient-specific calibration. Although the non-patient-specific calibration was less favorable than the air-fat-muscle calibration, it also highly correlated to phantom calibration (R^2=0.94) [19]. This suggests that air-fat-muscle and non-patient-specific calibration might be good alternatives for phantom calibration, which could allow fracture risk assessment based on FE modeling without the effort of scanning a calibration phantom along with the patient. Additionally, these methods have the potential to enable the use of phantomless CT scans retrospectively in order to validate the FE models developed for fracture risk assessment.

However, the study by Eggermont et al. only tested the air-fat-muscle and non-patient-specific calibration methods on well-protocolized CT scans [19]. It is known that changes in scan protocol can result in different HU, BMD and FE failure loads [28-34]. Depending on certain patient properties such as size, age or body part(s) that are to be scanned, deviations in scanning protocols are frequently made, e.g., to alter the radiation dose. It is known that deviations in scanning protocols can affect the HU, potentially resulting in different BMD and calculated failure loads [28-39]. For example, differences in reconstruction kernel can significantly influence BMD measurements and failure load calculated by FE models using phantom calibration [28, 29, 31].

For air-fat-muscle calibration and non-patient-specific calibration, the influence of differences in scan protocol has not been studied yet. Evaluation of the influence of commonly changed CT protocols on FE models when using phantomless calibration instead of phantom calibration is necessary to find out if these calibrations are equally feasible to use as phantom calibration. Therefore, the aim of this study is to investigate if air-fat-muscle calibration or non-patient-specific calibration can be used instead of phantom calibration when different CT protocols are used. Differences in effect of CT protocols on HU, BMD and FE failure loads between phantom and the two phantomless calibration methods were studied.

METHODS

CT scanning

Five fresh frozen human lower limbs (3 female, 2 male; mean age 70.4, range 63-77; 2 left, 3 right) were obtained from the Anatomy Department of the Radboud university medical center (Radboudumc, https://www.radboudumc.nl/en/research/departments/anatomy) according to the Dutch Body Donation Program for Science and Education [40].

All lower limbs were scanned on a Philips Brilliance Big Bore (Philips Medical Systems, Eindhoven, The Netherlands) in supine human orientation (patella facing up, proximal end

first) on top of a solid calibration phantom (Image analysis, Columbia, KY). This calibration phantom contained four rods with known calciumhydroxyapatite (CaHA) concentrations (0, 50, 100, and 200 mg/cm^3) and can be used to calibrate the HU of the scan to CaHA density as a measure of BMD (in mg/cm^3) [11, 12, 19, 29, 41-43]. All limbs were scanned using the standard protocol that we use for FE modeling [11, 12, 19] and with seven additional commonly used protocols with variations on this standard protocol which were selected based on the literature [31, 34, 35, 37-39], as well as our observance in our previous patient studies [12, 29] (Table 1). Eight different scans per leg and 40 scans in total were made. For all protocols, the pitch was set at 0.813.

Table 1. Overview of the different scan protocols and the corresponding CT settings. Standard protocol covers the CT settings that are standard for scans made at the Radiotherapy department of the Radboudumc. All other protocols deviate one setting from the standard protocol.
*For the 140 kV scan, the scanner software did not allow the standard tube current-time of 220 mAs. Hence, we used the highest value possible, which was 218 mAs.

	Value in standard protocol	Value in deviating protocol
Current-time (mAs)	220	163
Tube voltage (kV)	120	140*
Slice thickness (mm)	3	1
Rotation time (s)	1	0.75
FOV (mm)	480	600
Reconstruction kernel	B (for bone imaging)	C (for sharper images)
Reconstruction algorithm	standard	iDose4

FE modeling

Subject-specific FE models were created for each femur based on the CT scan made with the standard protocol, as described before [11, 12, 42]. Subject-specific femoral geometry was obtained from the CT scan made with the standard protocol by selecting the voxels containing femoral tissue (Mimics 14.0, Materialise, Leuven, Belgium). Femoral geometries were modeled to a solid mesh consisting of tetrahedral elements (Patran 2011, MSC Software Corporation, Santa Ana, CA, USA). The femoral geometry was registered to all other CT scans (elastix [45, 46]), followed by an accuracy check. The HU of each element was calibrated to BMD (in mg/cm^3) using the air-fat-muscle, non-patient-specific, and phantom calibration [19]. All the mentions of BMD in the study indicate "calculated BMD". For phantom calibration, nine diaphyseal slices were selected. The HU within the rods of the phantom in these slices were linearly correlated to the known CaHA concentrations to obtain the phantom calibration function. For air-fat-muscle calibration, a square region of interest was defined around the leg, including approximately 1 cm of air on each side of the leg, on the same nine diaphyseal slices that were used for the phantom calibration (Fig 1 left). The peaks of air, fat, and muscle tissue were extracted from a combined histogram of all HUs in the region of interest (Fig 1 right). Subsequently, to obtain the air-fat-muscle calibration function, the extracted HU peaks were linearly fitted to the reference BMD values (-838 mg/cm^3 for air, -86 mg/cm^3 for fat and 35 mg/cm^3 for muscle respectively) for each subject [19]. For the non-patient-specific calibration we used the same function as determined previously (BMD = 0.82 x HU–4.2 [19]), which was determined by averaging calibration functions of 26 patients. This calibration function was then applied to each of the 5 subjects.

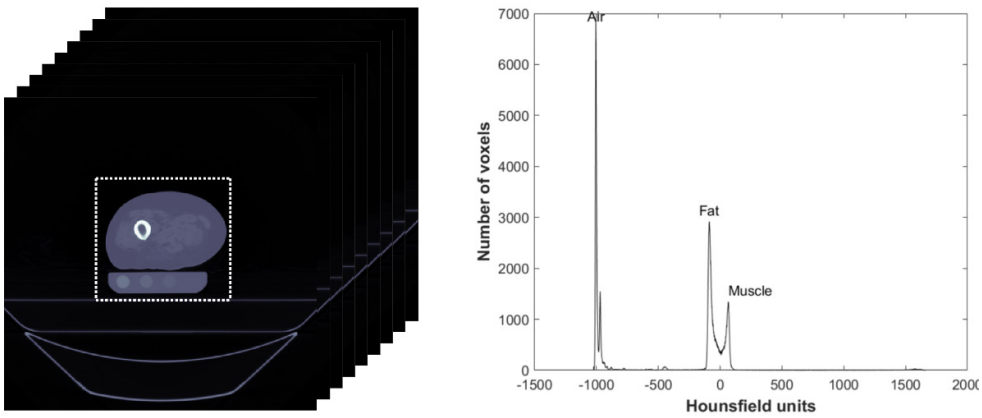


Fig 1. An example of the selected region of interest (left) and the histogram of the Hounsfield units within the region of interest with the peaks for air, fat, and muscle (right).

The BMD values (in mg/cm^3) were eventually converted to ash densities (ρ_{ash} = 0.000887 x BMD + 0.0633) which were converted to non-linear isotropic material properties [47]. Some material properties were different between trabecular and cortical bone (threshold BMD = 250 mg/ cm^3), using elastic modulus (MPa) = 14900 x $\rho_{ash}^{1.86}$; ultimate stress (MPa) = 102 x $\rho_{ash}^{1.80}$; plastic strain (mm/mm) at initial plastic phase = 0.00189 + 0.0241 x ρ_{ash} for BMD ≤ 250 mg/cm^3 and 0.0184 – 0.0100 x ρ_{ash} for BMD > 250 mg/cm^3; plastic modulus (MPa) of strain softening phase = –2080 x $\rho_{ash}^{1.45}$ for BMD ≤ 250 mg/cm^3 and – 1000 for BMD > 250 mg/cm^3; and stress (MPa) at the indefinite perfectly plastic phase = 43.1 x $\rho_{ash}^{1.81}$. The model was distally fixated by two bundles of high-stiffness springs. A cup representing the acetabulum was simulated on top of the femoral head and applied a displacement-driven load in the axial direction with increments of 0.1 mm (Fig 2). FE simulations were carried out using MSC. MARC (v2013.1, MSC Software Corporation, Santa Ana, CA, USA). Force-displacement (FD) curves were generated based on the displacement of the cup and the contact forces in the axial direction. It was assumed that a fracture occurred when the maximum total reaction force at the interface of the cup and femoral head was reached in the non-linear part of the load-displacement curve, which was defined as the failure load. For every femur, only BMD and subsequent mechanical properties varied between protocols, whereas all other model properties were left unchanged.

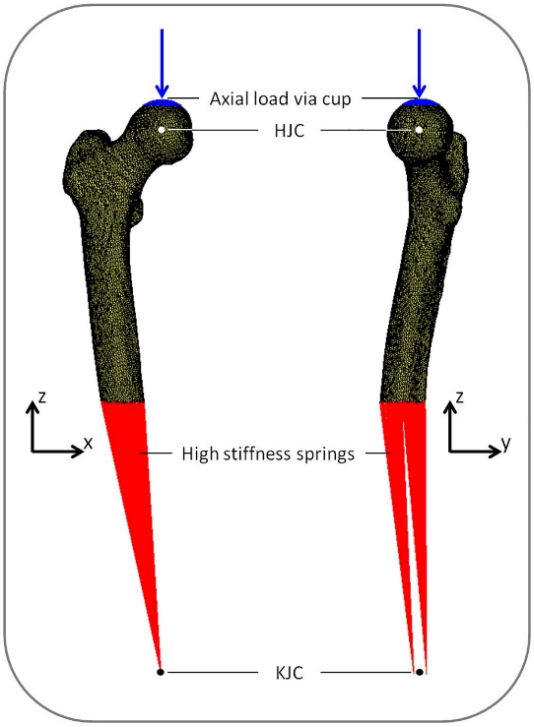


Fig 2. Boundary conditions of the finite element model. The model was distally fixated by two high stiffness springs, and the load was applied through a cup located on the head of the femur. The figure is with permission adapted from Eggermont et al., Bone, 2020 [11].

Outcome measures

To analyze HU and BMD, a cortical and a trabecular region of interest (ROI) were selected by selecting elements along the cortex of the femoral shaft and center of the femoral head, respectively (Fig 3). For each ROI, HUs were obtained, which were subsequently calibrated to BMDs using the phantom calibration [11, 12, 19, 29], air-fat-muscle calibration [19], and non-patient-specific calibration [19].

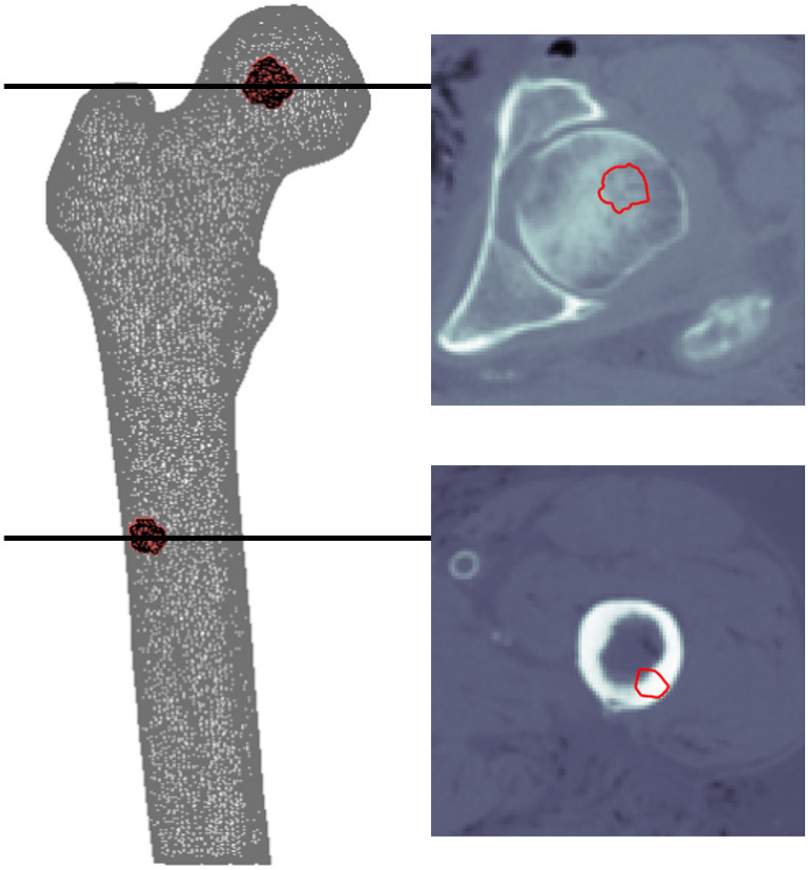


Fig 3. An example of the trabecular (upper) and cortical (lower) regions of interest including axial views.

Additionally, we calculated FE failure loads based on phantom, air-fat-muscle, and non-patient-specific calibration. For every protocol and calibration method, the HU and BMD within the cortical and trabecular ROIs, as well as the failure loads, were analyzed and compared. Basic statistics, i.e. averages and correlations were calculated.

RESULTS

Hounsfield unit

Mean HU in trabecular and cortical ROI decreased with 10% and 9.2% when changing the tube voltage from 120 to 140 kV, respectively. All other protocol variations had little effect, i.e. on average less than 2%, on the HU (Fig 4).

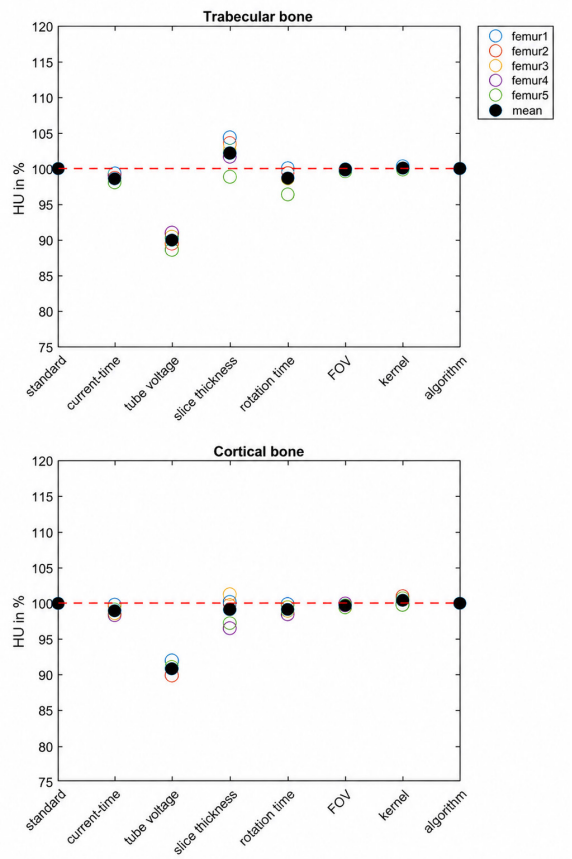


Fig 4. HU of the trabecular (upper figure) and cortical (lower figure) ROIs. For each femur, the HUs are displayed relative to the standard protocol. The red dashed line indicates 100%.

Bone mineral density

Generally, BMDs were 3.3% and 3.9% higher after air-fat-muscle and non-patient-specific calibrations compared to phantom calibration, respectively. Considering the standard protocol, the average absolute BMDs of cortical bone were 867 mg/cm^3 (range 666 to 1084 mg/cm^3) for phantom calibration, 895 mg/cm^3 (range 743 to 1089 mg/cm^3) for air-fat-muscle calibration, and 912 mg/cm^3 (range 701 to 1126 mg/cm^3) for non-patient-specific calibration. For trabecular bone, these values were 246 mg/cm^3 (range 214 to 269 mg/cm^3), 254 mg/cm^3 (range 208 to 281 mg/cm^3) and 267 mg/cm^3 (range 231 to 296 mg/cm^3), respectively. When phantom calibration was used, the relative BMD showed very little variation between the different protocols (Fig 5). For both air-fat-muscle and non-patient-specifc calibration, the BMD after changing tube voltage from 120 to 140 kV was relatively lower: for trabecular bone, the differences with the standard protocol were 9.8% and 10.1%, respectively, and for cortical bone, the differences with the standard protocol were 8.9% and 9.3% respectively (Fig 5).

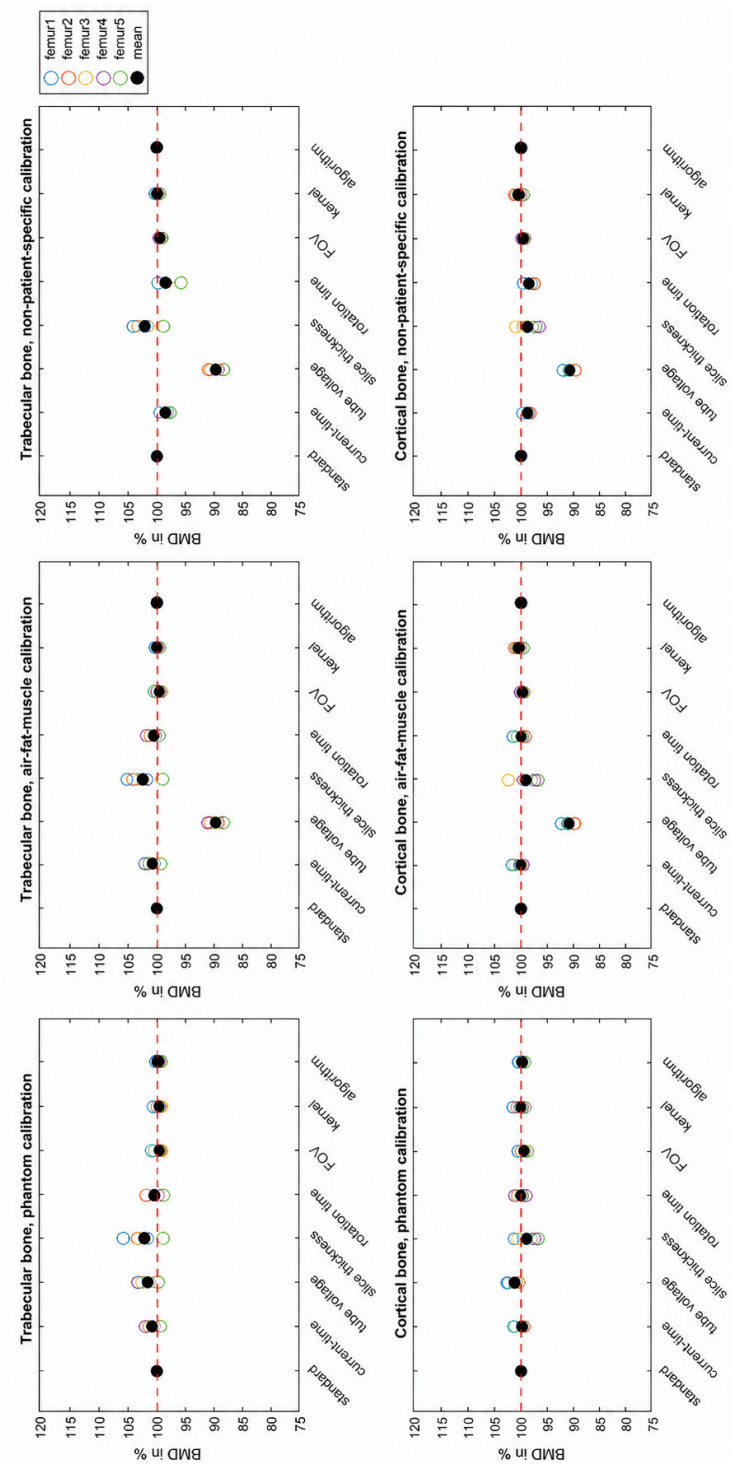


Fig 5. BMD of the trabecular (upper half) and cortical bone (lower half) for phantom (left), air-fat-muscle (middle), and non-patient-specific (right) calibration per protocol. For each femur, the BMD values are displayed relative to the standard protocol. The red dashed line indicates 100%.

Failure load

The failure load of the separate femurs obtained from FE analysis showed strong correlation between calibration methods (Fig 6). Only femur 5 showed a remarkable overestimation in failure load after air-fat-muscle calibration compared to phantom calibration. Higher correlations between phantom and air-fat-muscle calibration were obtained when excluding femur 5 (S1 Appendix).

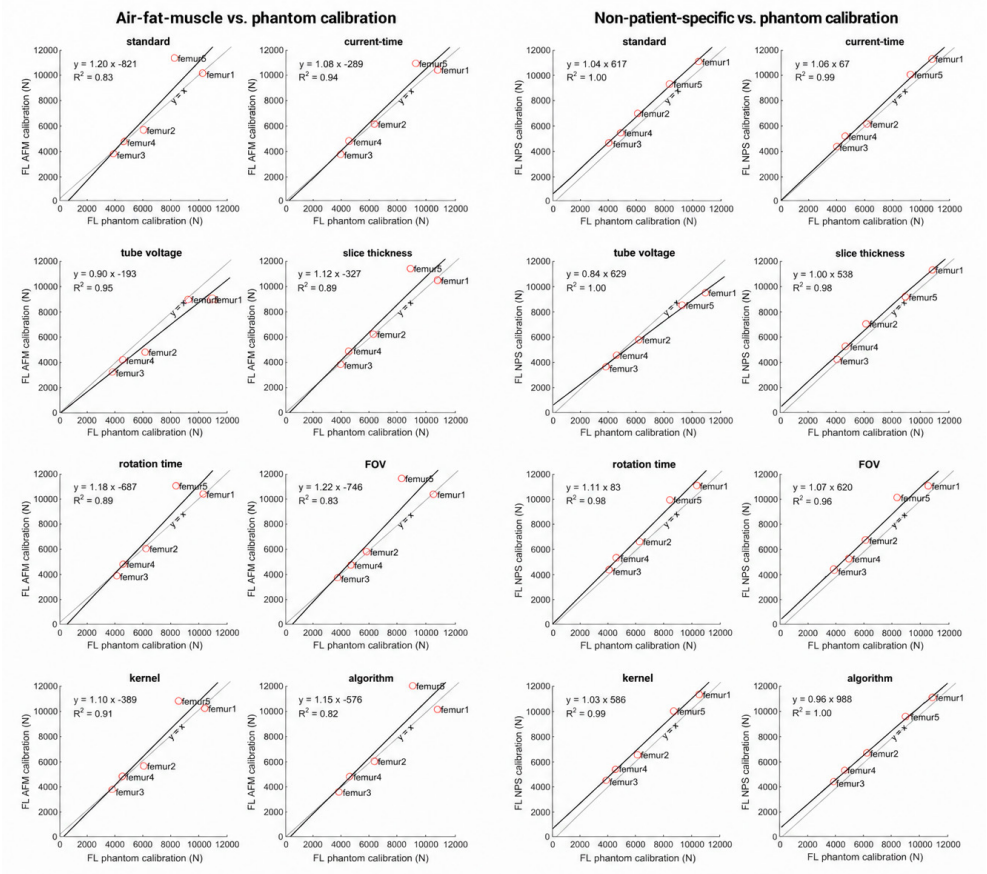


Fig 6. Correlation between failure loads after phantom and air-fat-muscle calibration and phantom and non-patient-specific calibration. Solid lines show the best-fit lines for each protocol.

The average failure load for the standard protocol was 6632 N (range 3967 to 10218 N) after phantom calibration, 7110 N (range 3759 to 11297 N) after air-fat-muscle calibration and 7505N (range 4749N to 11136 N) after non-patient-specific calibration. The difference in failure load relative to the standard protocol varied between femurs (Fig 7). Maximum mean

failure load variations from the standard protocol were 4.7% for phantom calibration, 13.8% for air-fat-muscle calibration, and 13.4% for non-patient-specific calibration. Most noticable were the relatively large decreases in failure load with 13.8% and 13.4% when changing tube voltage from 120 to 140 kV after air-fat-muscle and non-patient-specific calibration, respectively.

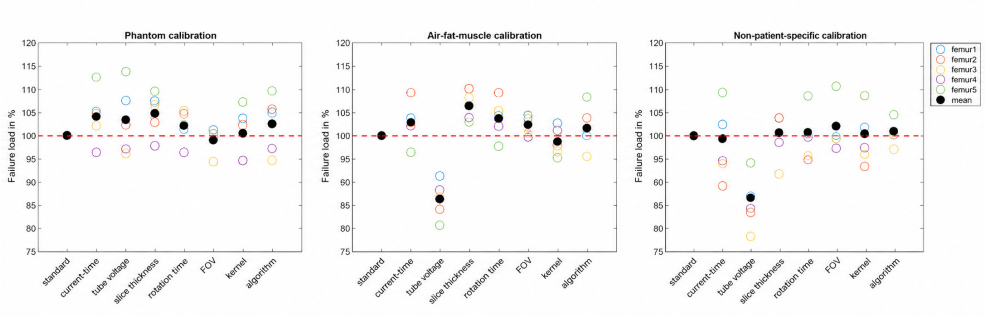


Fig 7. Failure load per protocol, per calibration method. For each femur, the failure loads are displayed relative to the standard protocol. The red dashed line indicates 100%.

DISCUSSION

The aim of this study was to investigate the differences in effect of common variations in CT scan protocols on HU, BMD and FE failure loads between phantom and air-fat-muscle calibration and phantom and non-patient-specific calibration, to determine whether air-fat-muscle and non-patient-specific calibration perform similar to phantom calibration. Most of the protocols had little effect on HU. Only the mean HU in trabecular and cortical ROI decreased when changing the tube voltage from 120 to 140 kV. Absolute BMDs were higher after air-fat-muscle calibration compared to phantom calibration and were even higher after non-patient-specific calibration. The relative BMD showed very little variation for phantom calibration. The difference in failure load between the standard protocols for phantom and air-fat-muscle as well as non-patient-specific calibration was very small and both phantomless methods showed a strong correlation with phantom calibration. Per femur, the effect of the deviating protocols on the failure load varied considerably. On average, changing tube voltage from 120 to 140 kV resulted in a relatively large decrease in failure load for the air-fat-muscle as well as the non-patient-specific calibration methods in comparison to the phantom calibration. The other variations did not noticeably deviate in failure load in comparison to the standard protocol.

The strong correlation between phantom and air-fat-muscle calibration methods allows their use to be interchangeable in case the air, fat, and muscle peaks are clearly detectable in the HU histogram. Also phantom and non-patient-specific calibration were strongly correlated. Hence non-patient-specific calibration is also useable, although in a previous study including 67 femurs

using the standard protocol, a small and significant difference in failure load between phantom and non-patient-specific calibration was present [19].

One must be critical about the implications that minor differences can have on the eventual failure risk assessment in patients. The treatment that a patient receives is partially based on a certain fracture risk threshold: a risk below the threshold requires conservative radiotherapy and a risk above the threshold requires stabilizing surgery [11]. When the fracture risk is close to the threshold, a minor failure load difference between calibration methods can influence the treatment plan. However, besides the calculated fracture risk, clinicians take other patient factors such as pain level, clinical condition, and life expectancy into consideration as well to come to a thoughtful treatment plan. As for the fracture risk assessment based on retrospective CT scans and comparing this with actual fracture occurrence, air-fat-muscle as well as non-patient-specific calibration seem to be a reliable alternatives, not only for metastatic femurs, but probably also for CT scans of patients with osteoporosis.

An exception to the strong correlation between the phantom and air-fat-muscle calibration methods was femur 5, which showed a much higher failure load after air-fat-muscle calibration (Fig 8). Upon inspection of the scans of this femur, accumulation of fat in muscle tissue was observed, causing a downgrade of muscle quality and density (Fig 8). Therefore, one should check whether the fat and muscle peaks are clearly detectable prior to using air-fat-muscle calibration.

The influence of the different scan protocols on the failure load with respect to the standard protocol varied per femur. Maximum mean failure load differences were 4.7% for phantom calibration, 13.8% for air-fat-muscle calibration and 13.4% for non-patient-specific calibration, implying that phantom calibration after scanning with deviating protocols returns more robust results than air-fat-muscle or non-patien-specific calibration.

Changing tube voltage showed the highest relative difference in failure load after air-fat-muscle and non-patient-specific calibration compared to the standard protocol. Giambini et al. investigated the influence of varying tube voltage values on the HU of bone [31]. They showed that increasing tube voltage caused a drop in HU. On the other hand, they found little variation in cortical BMD between the two tube voltage values. These findings concur with the results of the phantom calibrated scans found in the present study: the HU of the calibration phantom rods were relatively lower for the tube voltage protocol compared to the standard protocol (S1 table), but the BMD barely differed between protocols. Subsequently, no remarkable differences in failure load were observed. Apparently, in contrast with the air-fat-muscle and non-patient-specific calibration, the phantom calibration method is able to correct for any changes in tube voltage , thus nullifying the influence of the decreased HU on the calibration. For the air-fat-muscle calibration, this can be explained by the fact that HU are less influenced by tube voltage

differences when the electron density of the tissues gets below the electron density of water [48], and thus the effect of tube voltage deviations might be limited in fat and muscle tissue [49]. Hence, the calibration curve of air-fat-muscle is practically similar to the standard protocol, even though the HU of bone will be lower because of the higher tube voltage . On the contrary, as the densities of the rods in the calibration phantom are higher than water, the effect of tube voltage deviations is more noticeable, resulting in a different calibration curve that is able to correct for the change in tube voltage in bone HU. Additionally, it is evident that non-patient-specific calibration cannot correct for any differences caused by changes in tube voltage . Hence, it is advised to use the same tube voltage for the scans that are calibrated with air-fat-muscle or non-patient-specific methods.

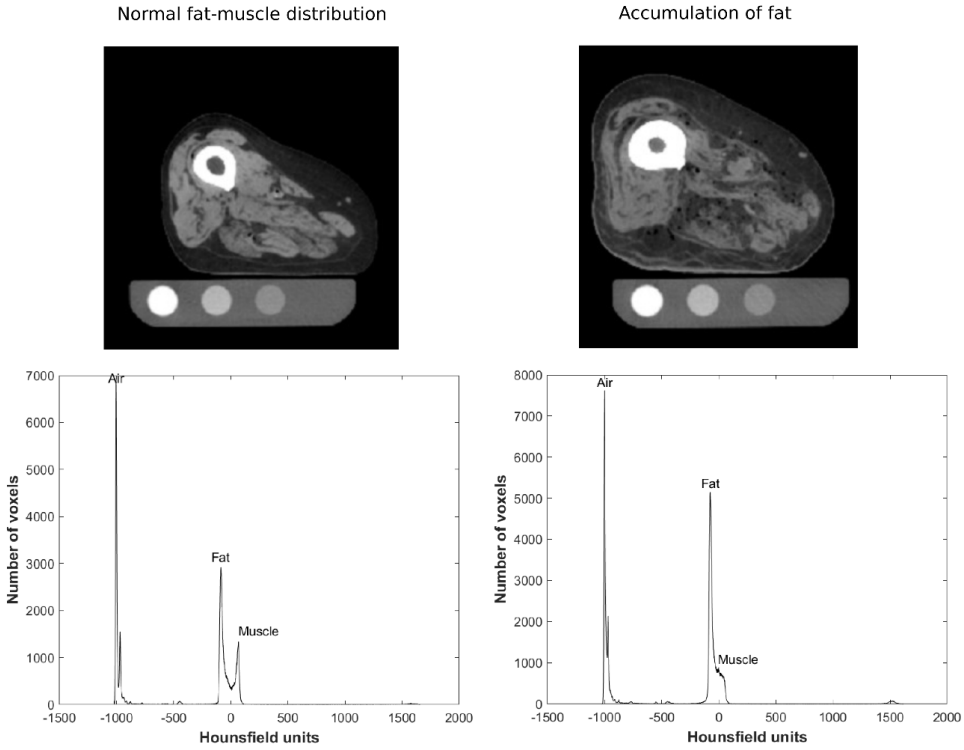


Fig 8. Slices of the same diaphyseal region from the CT scans of femur 4 (left) and femur 5 (right) with the corresponding HU histograms. Femur 5 shows accumulation of fat in the muscle tissue.

Additionally, one should notice that reconstruction methods differ between different types of CT scanners. Following a standard protocol for one scanner might therefore not yield the same BMD on another scanner, potentially resulting in a different failure risk calculation [28, 32]. Eggermont et al. found significant differences in HU (average 7±2%) and BMD (max 6±1%) between four different scanners using phantom calibration, whereas failure load did only

significantly differ in one scan as opposed to the other three (max 17±5%) [29]. Birnbaum et al. support this finding and observed a significant 9-10 HU difference between scanners in renal cyst tissue [36]. Non-patient-specific calibration will not be able to correct for any differences between CT scanners. Future research should point out how different scanners affect air-fat-muscle calibration and resulting failure load.

Additionally, in our FE simulation only one axial loading condition was applied. However, our FE model with the relatively simple axial loading condition showed an improvement of fracture risk assessments compared to methods used in clinical guidelines in several previous studies [11, 12]. Although multiple loading conditions may lead to more information about the fracture risk, using one loading condition was very useful for clinical implementation of the FE model [50].

We acknowledge some limitations in this study. First of all, a relatively small number of cadaveric legs was used in the present study. Due to the small sample size we were not able to conduct any reliable statistical analysis. However, the correlation between failure load after phantom calibration and failure load after air-fat-muscle calibration was comparable with the data from a study by Eggermont et al. (S2 Appendix) in which 67 femurs were included [19]. Therefore, we think it is plausible that tube voltage significantly affects the HU, BMD, and failure load compared to the other protocol variations. Remarkably, the femurs calibrated with non-patient-specific calibration seemed to deviate more from the data of Eggermont et al. [19]. This might be explained by the fact that the femurs in the current study were scanned as single legs, whereas the femurs of Eggermont et al. were actual patients. When scanning a single leg, the HU will be higher because of less absorption of other tissue (due to a missing contralateral leg). As a result the calculated failure loads will be too high, for which non-patient-specific calibration apparently cannot correct, whereas air-fat-muscle calibration seems to be able to do this better. A second limitation relates to the effect of cryopreservation on subjects. It is known that freezing and subsequent thawing can influence the HU of tissue because of differences in water content [51, 52], which might have affected the subjects in this study. However, as we compare three calibration methods and use the same material for all methods, the effect of freezing and thawing on the conclusions of this study are expected to be negligible. Moreover, conducting a similar study with living subjects would be unethical due to the high radiation doses. A third limitation was the lack of data on how the subjects passed away and if any diseases played a role. These factors could have had an influence on the tissues of the subjects, and subsequently on the BMD and failure load after air-fat-muscle calibration. Femur 5 is such an example. Finally, we only tested a few protocol variations. Since we could not investigate the effect of all possible variations in CT scan protocols and we are aiming at increasing the clinical implementation of femoral fracture risk assessment, we selected the protocols based on the literature, as well as protocol violations that we came across doing our clinical patient

study. These protocol variations were subsequently examined in this study. We only tested the protocols on one CT scanner. In the future, it should be investigated whether the air-fat-muscle calibration varies compared to phantom calibration when using different CT scanners.

In conclusion, while we observed significant correlations between air-fat-muscle calibration and phantom calibration as well as between non-patient-specific calibration and phantom calibration, our sample size was too small to prove that either of these calibration approaches was superior. In addition, our approaches for air-fat-muscle and non-patient-specific calibration should not be used for variations in tube voltage. These should be corrected for. The influence of tissue abnormalities on air-fat-muscle calibration compared to phantom calibration should be studied explicitly to prevent discrepancies in fracture risk assessment. Further studies are necessary to test whether air-fat-muscle or non-patient-specific calibration could replace phantom calibration in case of different scanning protocols and different CT scanners.

REFERENCES

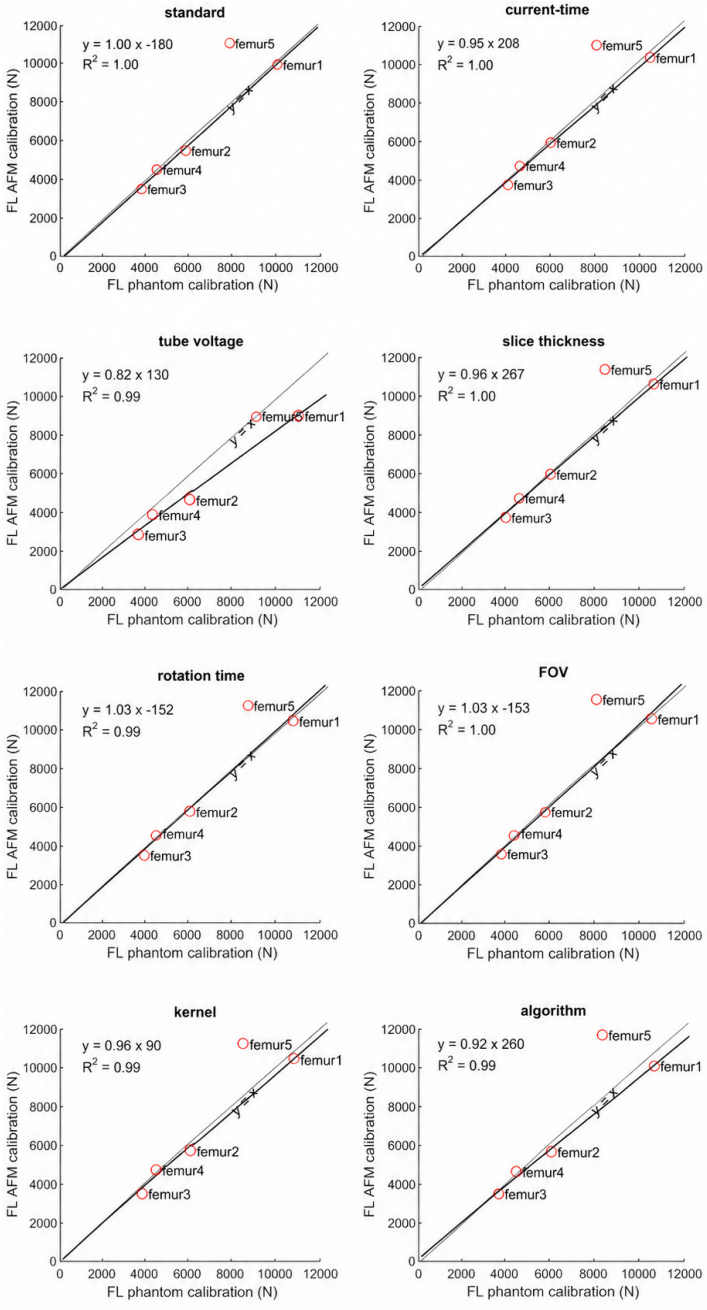
1. Johnson SK, Knobf MT. Surgical interventions for cancer patients with impending or actual pathologic fractures. Orthopaedic Nursing. 2008;27(3):160-71.
2. Ottanelli S. Prevention and treatment of bone fragility in cancer patient. Clin Cases Miner Bone Metab. 2015;12(2):116-29.
3. Van der Linden YM, Dijkstra PD, Kroon HM, Lok JJ, Noordijk EM, Leer JW, et al. Comparative analysis of risk factors for pathological fracture with femoral metastases. J Bone Joint Surg Br. 2004;86(4):566-73.
4. Ulaner GA, Zindman AM, Zheng J, Kim TW, Healey JH. FDG PET/CT Assesses the Risk of Femoral Pathological Fractures in Patients With Metastatic Breast Cancer. Clin Nucl Med. 2017;42(4):264-70.
5. Mirels H. Metastatic disease in long bones. A proposed scoring system for diagnosing impending pathologic fractures. Clin Orthop Relat Res. 1989(249):256-64.
6. Damron TA, Morgan H, Prakash D, Grant W, Aronowitz J, Heiner J. Critical evaluation of Mirels' rating system for impending pathologic fractures. Clinical Orthopaedics and Related Research®. 2003;415:S201-S7.
7. Tatar Z, Soubrier M, Dillies AF, Verrelle P, Boisgard S, Lapeyre M. Assessment of the risk factors for impending fractures following radiotherapy for long bone metastases using CT scan-based virtual simulation: a retrospective study. Radiation oncology. 2014;9(1):227.
8. Shimoyama T, Katagiri H, Harada H, Murata H, Wasa J, Hosaka S, et al. Fracture after radiation therapy for femoral metastasis: incidence, timing and clinical features. Journal of Radiation Research. 2017;58(5):661-8.
9. van der Linden YM, Kroon HM, Dijkstra SP, Lok JJ, Noordijk EM, Leer JWH, et al. Simple radiographic parameter predicts fracturing in metastatic femoral bone lesions: results from a randomised trial. Radiotherapy and Oncology. 2003;69(1):21-31.
10. Van der Wal C, Eggermont F, Fiocco M, Kroon H, Ayu O, Slot A, et al. Axial cortical involvement of metastatic lesions to identify impending femoral fractures; a clinical validation study. Radiotherapy and Oncology. 2020;144:59-64.
11. Eggermont F, van der Wal G, Westhoff P, Laar A, de Jong M, Rozema T, et al. Patient-specific finite element computer models improve fracture risk assessments in cancer patients with femoral bone metastases compared to clinical guidelines. Bone. 2020;130:115101.
12. Eggermont F, Derikx LC, Verdonschot N, van der Geest ICM, de Jong MAA, Snyers A, et al. Can patient-specific finite element models better predict fractures in metastatic bone disease than experienced clinicians?: Towards computational modelling in daily clinical practice. Bone Joint Res. 2018;7(6):430-9.
13. Sternheim A, Giladi O, Gortzak Y, Drexler M, Salai M, Trabelsi N, et al. Pathological fracture risk assessment in patients with femoral metastases using CT-based finite element methods. A retrospective clinical study. Bone. 2018;110:215-20.
14. Goodheart JR, Cleary RJ, Damron TA, Mann KA. Simulating activities of daily living with finite element analysis improves fracture prediction for patients with metastatic femoral lesions. Journal of Orthopaedic Research. 2015;33(8):1226-34.
15. Rajapakse CS, Gupta N, Evans M, Alizai H, Shukurova M, Hong AL, et al. Influence of bone lesion location on femoral bone strength assessed by MRI-based finite-element modeling. Bone. 2019;122:209-17.

16. Reeves JM, Knowles NK, Athwal GS, Johnson JA. Methods for Post Hoc Quantitative Computed Tomography Bone Density Calibration: Phantom-Only and Regression. J Biomech Eng. 2018;140(9).
17. Boomsma MF, Slouwerhof I, van Dalen JA, Edens MA, Mueller D, Milles J, et al. Use of internal references for assessing CT density measurements of the pelvis as replacement for use of an external phantom. Skeletal Radiol. 2015;44(11):1597-602.
18. Habashy AH, Yan X, Brown JK, Xiong X, Kaste SC. Estimation of bone mineral density in children from diagnostic CT images: a comparison of methods with and without an internal calibration standard. Bone. 2011;48(5):1087-94.
19. Eggermont F, Verdonschot N, van der Linden Y, Tanck E. Calibration with or without phantom for fracture risk prediction in cancer patients with femoral bone metastases using CT-based finite element models. PLoS One. 2019;14(7):e0220564.
20. Lee DC, Hoffmann PF, Kopperdahl DL, Keaveny TM. Phantomless calibration of CT scans for measurement of BMD and bone strength—inter-operator reanalysis precision. Bone. 2017;103:325-33.
21. Boomsma MF, Slouwerhof I, van Dalen JA, Edens MA, Mueller D, Milles J, et al. Use of internal references for assessing CT density measurements of the pelvis as replacement for use of an external phantom. Skeletal radiology. 2015;44(11):1597-602.
22. Weaver AA, Beavers KM, Hightower RC, Lynch SK, Miller AN, Stitzel JD. Lumbar bone mineral density phantomless computed tomography measurements and correlation with age and fracture incidence. Traffic injury prevention. 2015;16(sup2):S153-S60.
23. Mueller DK, Kutscherenko A, Bartel H, Vlassenbroek A, Ourednicek P, Erckenbrecht J. Phantom-less QCT BMD system as screening tool for osteoporosis without additional radiation. European journal of radiology. 2011;79(3):375-81.
24. Boden SD, Goodenough DJ, Stockham CD, Jacobs E, Dina T, Allman RM. Precise measurement of vertebral bone density using computed tomography without the use of an external reference phantom. Journal of digital imaging. 1989;2(1):31-8.
25. Weaver AA, Beavers KM, Hightower RC, Lynch SK, Miller AN, Stitzel JD. Lumbar Bone Mineral Density Phantomless Computed Tomography Measurements and Correlation with Age and Fracture Incidence. Traffic Inj Prev. 2015;16 Suppl 2:S153-60.
26. Winsor C, Li X, Qasim M, Henak C, Pickhardt P, Ploeg H, et al. Evaluation of patient tissue selection methods for deriving equivalent density calibration for femoral bone quantitative CT analyses. Bone. 2021;143:115759.
27. Lee YH, Kim JJ, Jang IG. Patient-specific phantomless estimation of bone mineral density and its effects on finite element analysis results: a feasibility study. Computational and mathematical methods in medicine. 2019;2019.
28. Free J, Eggermont F, Derikx LC, van Leeuwen R, van der Linden YM, Jansen W, et al. The effect of different CT scanners, scan parameters and scanning setup on Hounsfield units and calibrated bone density: a phantom study. Biomedical Physics & Engineering Express. 2018;4(5):055013.
29. Eggermont F, Derikx LC, Free J, van Leeuwen R, van der Linden YM, Verdonschot N, et al. Effect of different CT scanners and settings on femoral failure loads calculated by finite element models. J Orthop Res. 2018.

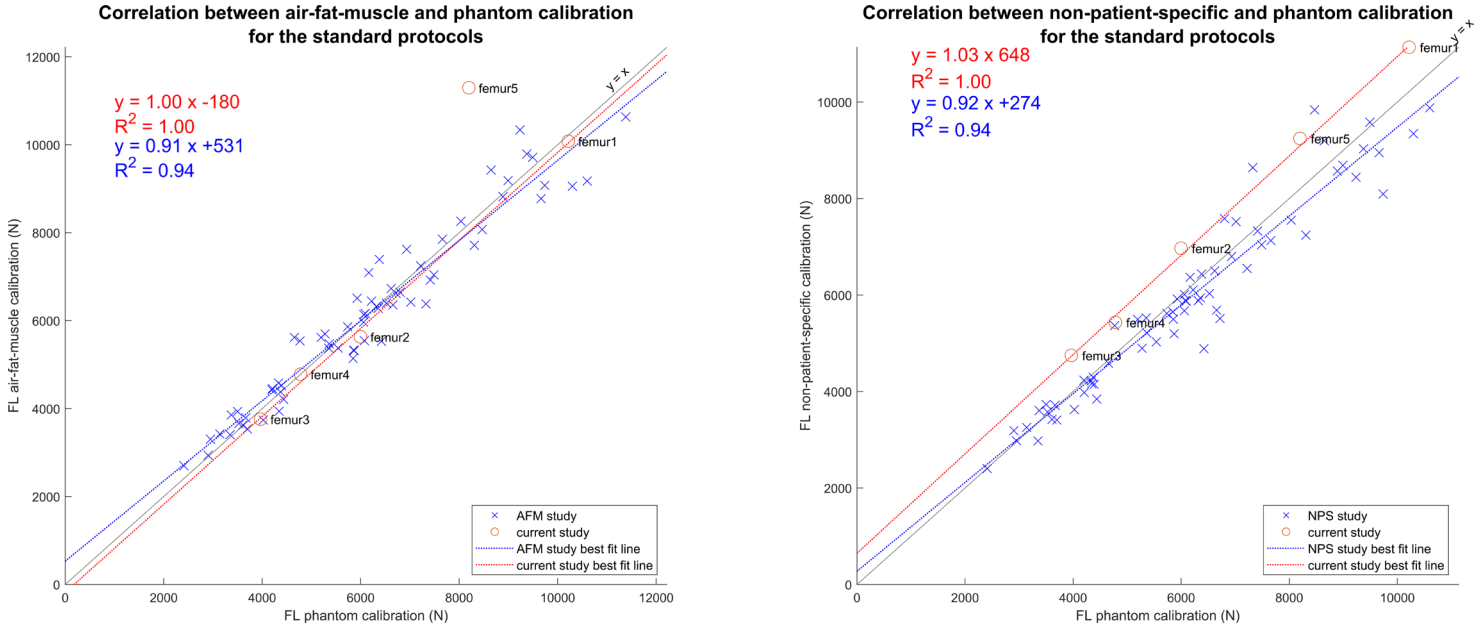
30. Carpenter RD, Saeed I, Bonaretti S, Schreck C, Keyak JH, Streeper T, et al. Inter-scanner differences in in vivo QCT measurements of the density and strength of the proximal femur remain after correction with anthropomorphic standardization phantoms. Med Eng Phys. 2014;36(10):1225-32.
31. Giambini H, Dragomir-Daescu D, Huddleston PM, Camp JJ, An KN, Nassr A. The Effect of Quantitative Computed Tomography Acquisition Protocols on Bone Mineral Density Estimation. J Biomech Eng. 2015;137(11):114502.
32. Knowles NK, Reeves JM, Ferreira LM. Quantitative Computed Tomography (QCT) derived Bone Mineral Density (BMD) in finite element studies: a review of the literature. J Exp Orthop. 2016;3(1):36.
33. Alshipli M, Kabir N. Effect of slice thickness on image noise and diagnostic content of single-source-dual energy computed tomography. Journal of Physics: Conf Series. 2017;851(1):012005.
34. Davis AT, Palmer AL, Pani S, Nisbet A. Assessment of the variation in CT scanner performance (image quality and Hounsfield units) with scan parameters, for image optimisation in radiotherapy treatment planning. Phys Med. 2018;45:59-64.
35. Kalra MK, Maher MM, Toth TL, Schmidt B, Westerman BL, Morgan HT, et al. Techniques and applications of automatic tube current modulation for CT. Radiology. 2004;233(3):649-57.
36. Birnbaum BA, Hindman N, Lee J, Babb JS. Multi-detector row CT attenuation measurements: assessment of intra- and interscanner variability with an anthropomorphic body CT phantom. Radiology. 2007;242(1):109-19.
37. Beeres M, Wichmann JL, Paul J, Mbalisike E, Elsabaie M, Vogl TJ, et al. CT chest and gantry rotation time: does the rotation time influence image quality? Acta Radiol. 2015;56(8):950-4.
38. Klink T, Regier M, van Stevendaal U, Grass M, Adam G, Begemann P. Accelerating image acquisition in 64-MDCT: the influence of scan parameters on image resolution and quality in a phantom study. Clin Imaging. 2012;36(4):334-44.
39. Sookpeng S, Martin CJ, Gentle DJ. Investigation of the influence of image reconstruction filter and scan parameters on operation of automatic tube current modulation systems for different CT scanners. Radiat Prot Dosimetry. 2015;163(4):521-30.
40. Wet op de lijkbezorging. 1991.
41. Derikx LC, Verdonschot N, Tanck E. Towards clinical application of biomechanical tools for the prediction of fracture risk in metastatic bone disease. Journal of biomechanics. 2015;48(5):761-6.
42. Derikx LC, van Aken JB, Janssen D, Snyers A, van der Linden YM, Verdonschot N, 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. The Journal of bone and joint surgery British volume. 2012;94(8):1135-42.
43. Derikx LC, Vis R, Meinders T, Verdonschot N, Tanck E. Implementation of asymmetric yielding in case-specific finite element models improves the prediction of femoral fractures. Computer methods in biomechanics and biomedical engineering. 2011;14(02):183-93.
44. Free J, Wang F, Williams N, Gundara JS, Staerkle RF, Hugh TJ, et al. Gallbladder mucosal lesions associated with high biliary amylase irrespective of pancreaticobiliary maljunction. ANZ J Surg. 2018;88(6):E517-E21.
45. Klein S, Staring M, Murphy K, Viergever MA, Pluim JP. elastix: a toolbox for intensity-based medical image registration. IEEE Trans Med Imaging. 2010;29(1):196-205.

46. Shamonin DP, Bron EE, Lelieveldt BP, Smits M, Klein S, Staring M, et al. Fast parallel image registration on CPU and GPU for diagnostic classification of Alzheimer's disease. Front Neuroinform. 2013;7:50.
47. Keyak JH, Kaneko TS, Tehranzadeh J, Skinner HB. Predicting proximal femoral strength using structural engineering models. Clin Orthop Relat Res. 2005(437):219-28.
48. Das IJ, Cheng CW, Cao M, Johnstone PA. Computed tomography imaging parameters for inhomogeneity correction in radiation treatment planning. J Med Phys. 2016;41(1):3-11.
49. Schneider W, Bortfeld T, Schlegel W. Correlation between CT numbers and tissue parameters needed for Monte Carlo simulations of clinical dose distributions. Phys Med Biol. 2000;45(2):459-78.
50. F. Eggermont PGW, I.C.M. van der Geest, N. Verdonschot, S. Ligthert, Y.M. van der Linden, E. Tanck. Implementation of the BOne Strength score for patients with bone metastases in the femur: first results and experiences. NED TIJDSCHR ONCOL. 2021;18:4-11.
51. Pech P, Bergstrom K, Rauschning W, Haughton VM. Attenuation values, volume changes and artifacts in tissue due to freezing. Acta Radiol. 1987;28(6):779-82.
52. Hemmingsson A, Johansson A, Rauschning W. Attenuation in human muscle and fat tissue in vivo and in vitro. Acta Radiol Diagn (Stockh). 1982;23(2):149-51.

APPENDIX



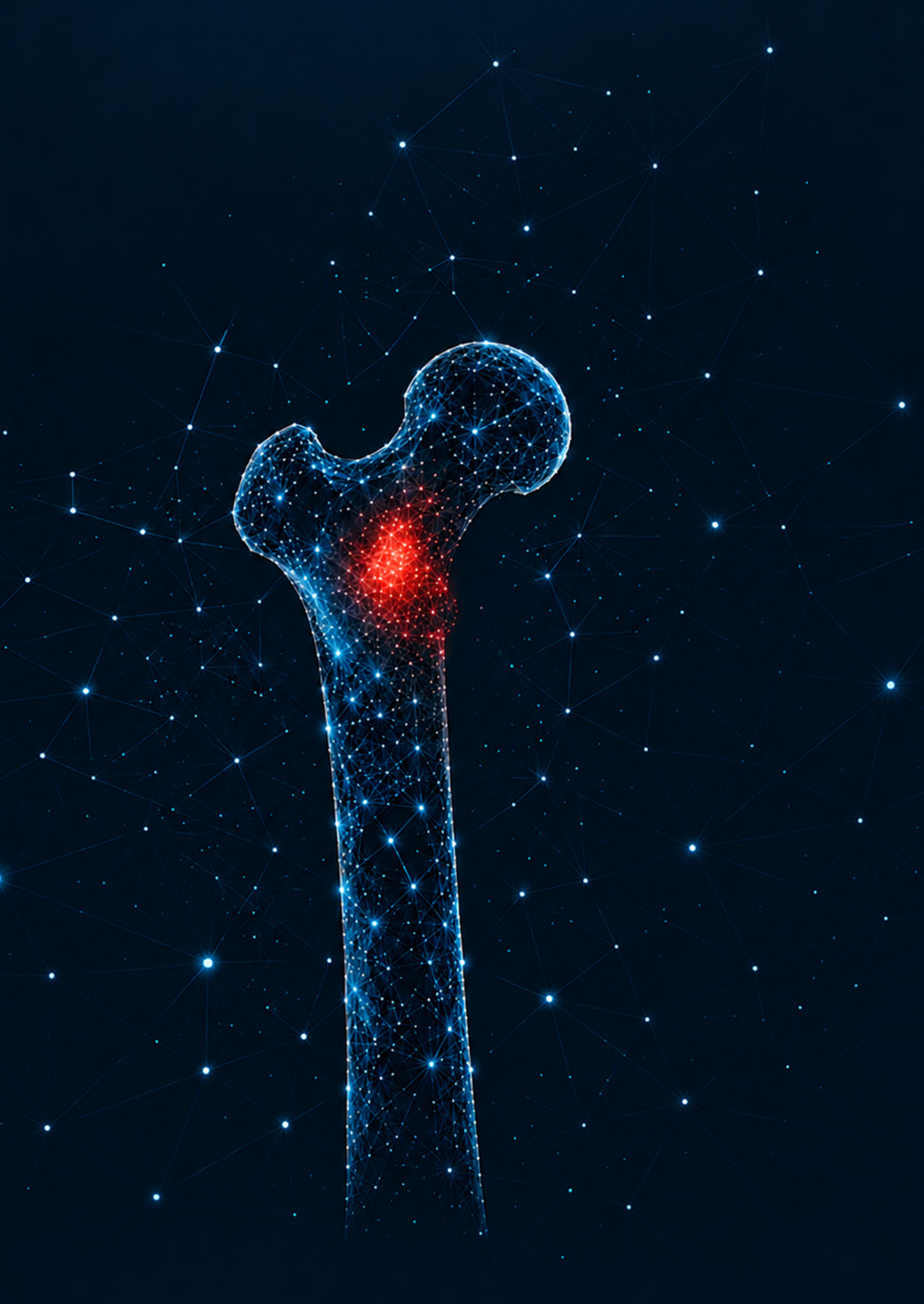
S1. Correlation between failure loads after phantom and air-fat-muscle calibration. Solid lines show the best-fit lines for each protocol in which femur 5 was excluded due to the accumulation of fat.



S2. Correlation between the failure load of the femurs of the previous study (Eggermont et al., PLOS ONE, 2019; blue line, n=67) compared to the correlation of the current study, left for air-fat-muscle and right for non-patient-specific calibration. This figure shows that in case of air-fat-muscle calibration four of the five femurs are within the range of the patients studied by Eggermont et al. (red line, n=4). Only the femur with the deviating fat and muscle balance is deviating. Additionally, it can be seen that the correlations between phantom and air-fat-muscle calibration are comparable when excluding the deviating femur. The femurs calibrated with non-patient-specific calibration seem to deviate more from the femurs of the previous study. This might be explained by the fact that the femurs in the current study are scanned as single legs, whereas the femurs of the previous study were actual patients. When scanning a single leg, the HU will be higher because of less absorption of other tissue (due to a missing contralateral leg). As a result, the failure loads will be too high, for which non-patient-specific calibration apparently cannot correct, whereas air-fat-muscle calibration seems to be able to do this better.

S1 Table. mean HUs per protocol for each calibration phantom rod.

	Mean HU			
	Calibration phantom rods			
Protocol	**0 mg/cm^3**	**50 mg/cm^3**	**100 mg/cm^3**	**200 mg/cm^3**
standard	16.51	82.146	145.552	269.184
current-time	12.078	76.94	140.13	262.936
tube voltage	12.194	71.132	128.382	239.126
slice thickness	16.238	81.856	145.652	268.768
rotation time	12.346	77.034	140.104	262.74
FOV	15.816	81.352	144.74	268.04
kernel	16.456	82.07	145.524	269.25
algorithm	16.672	82.22	145.82	269.364



Chapter 6

General Discussion

GENERAL DISCUSSION

This thesis focused on improvement of fracture risk assessment in patients with femoral bone metastases by advancing the clinical applicability of the finite element (FE) modeling workflow. While traditional tools like Mirels' score or axial cortical involvement have been widely used, their limited specificity and reliance on subjective interpretation restrict their clinical utility [1, 2]. To address this, the research explored key methodological developments within the FE modeling workflow, which forms the basis of the BOne Strength (BOS) score for estimating patient-specific fracture risk, such as the evaluation of deep learning for lesion segmentation (Chapter 2), the variation and more realistic adaptation of lesion-specific material properties (Chapters 3 and 4), and the assessment of phantomless CT scan calibration (Chapter 5). These studies aim to investigate the feasibility of these approaches and their potential to make the BOS score more robust, reproducible, and accessible in both prospective and retrospective clinical settings. Rather than providing ready-to-use clinical solutions, these studies revealed both the possibilities and the substantial limitations of current approaches, particularly in relation to segmentation accuracy and lesion-specific modeling.

Manual and AI-based segmentation

Accurate segmentation of metastatic lesions is an essential step for patient-specific fracture risk assessment using FE models, especially in patients with osteoblastic or mixed lesions. In contrast to osteolytic lesions, which are typically characterized by a decreased bone density and can be modeled using the normal bone density distribution, osteoblastic and mixed lesions require segmentation to enable proper adjustment of their material properties. However, manual segmentation, while currently considered the standard approach in clinical research, presents significant challenges and limitations. As shown in this thesis (Chapter 2) and supported by earlier studies, manual segmentation is time-consuming [3, 4] and prone to substantial inter- and intra-operator variability, especially in cases involving small, diffuse, or ambiguous lesions. This variability also challenges the reliability of manual annotations as ground truth data for training AI models. Furthermore, these inconsistencies compromise the reproducibility of FE-based fracture risk assessments and may weaken the objectivity of clinical decisions in patients with osteoblastic or mixed lesions. The implication is that without consistent segmentation, running a reliable fracture risk assessment for every patient with osteoblastic or diffuse bone metastasis becomes problematic, as different individuals who do the segmentations might produce different lesion maps leading to different risk assessments. Similar inconsistencies in lesion segmentation have also been shown to affect radiotherapy planning, where variations between clinicians can result in differences in treatment [5].

To address these challenges, artificial intelligence (AI), particularly deep learning, is recognized as a promising solution for lesion segmentation in medical imaging [6]. Convolutional neural

networks (CNNs), and more recently self-configuring architectures such as nnU-Net, have demonstrated strong performance in automatically segmentation of anatomical structures and tumors across a wide range of imaging modalities [7]. These AI models learn complex visual patterns from annotated datasets.

However, as indicated earlier, generating high-quality annotations for metastatic bone lesions is particularly challenging compared to other domains such as lung or breast imaging [8], where lesion boundaries are often more distinct and manual segmentations yield high inter-operator agreement. In contrast, bone metastases are harder to segment consistently, as also demonstrated in Chapter 2, which complicates both manual annotation and algorithm training. Even with advanced imaging, Gerlich et al. [9] evaluated interobserver agreement in the segmentation of bone metastases and reported conformity indices of 0.46 on CT, 0.54 on CT combined with MRI, and 0.56 on MRI, indicating considerable variability even among experienced clinicians.

These results are comparable to our findings, where mean inter-operator Dice coefficients ranged from 0.50 to 0.54 and intra-operator values from 0.56 to 0.71, highlighting how unclear lesion borders reduce segmentation consistency. This affects the ease with which AI models can generalize to new cases [10]. These challenges stem from imaging contrast variability, irregular mineralization, and the complex integration of lesions within bone structure [9]. Recent evidence specific to femoral osteolytic tumors reports nnU-Net test Dice scores of 0.67 and 0.68 against two expert radiologists, close to the radiologists' own inter-observer Dice (0.73), suggesting that automatic segmentation can approach expert performance in this indication [11].

In this thesis, the AI models did not consistently overcome these issues. While they were able to produce acceptable segmentations in many cases, their performance was highly variable, and in some patients the errors were large enough to substantially influence the predicted fracture risk. In practice, this means that an AI model might correctly segment most metastases but occasionally misclassify lesion boundaries or normal bone as lesion in challenging cases, which in turn could affect the calculated fracture risk for those individual patients. It is important to note that these AI models were trained on manual segmentations that showed limited inter- and intra-operator reliability, as demonstrated in Chapter 2. Consequently, the observed variability in AI performance should not be interpreted in isolation, since part of the deviation reflects uncertainty in the ground truth annotations rather than intrinsic deficiencies of the algorithm itself.

In skeletal oncology, where metastases can vary significantly in appearance and density, AI-based segmentation offers a major advantage by providing fast, reproducible, and observer-independent lesion contours. However, the findings of this thesis indicate that current AI frameworks, including nnU-Net, still fall short of providing robust output for every patient.

Chapters 3 and 4 showed that inaccuracies in lesion segmentation may have limited impact on fracture risk assessment at a group level. However, for individual patients, which is the primary focus of the BOS score, even small segmentation errors can lead to substantial deviations in the estimated fracture risk. Thus, while our AI-assisted approach improved the segmentation efficacy and standardization for large cohorts, achieving consistent accuracy at the individual level remains challenging.

Additionally, recent autonomous-FE findings show that a high Dice coefficient in lesion segmentation does not guarantee small FE-strain error [12]. Manual and automatic masks agreed in most regions but not all, and small boundary differences can still shift patient-level risk indices [12]. In the context of achieving a fracture risk assessment for every patient with bone metastasis, the segmentation method must be exceptionally robust for each individual case, a requirement that is not yet fully met. This highlights a limitation of current supervised learning models, which rely heavily on manual annotations as training input.

Given the variability and uncertainty in those annotations, alternative AI strategies may be needed. One potential solution is to incorporate multi-modality imaging or advanced model uncertainty estimates. Multi-modality imaging, for instance combining CT and MRI, may reduce ambiguity in lesion boundaries and improve the consistency of training data, which could enhance segmentation robustness [13]. Incorporating uncertainty estimation, as demonstrated in probabilistic U-Net variants, allows the model to produce several alternative segmentation hypotheses and their likelihoods, explicitly flagging ambiguous areas for targeted clinical review and follow-up actions [14].

Once segmentation can reliably produce standardized lesion descriptors such as volume, anatomical location, cortical proximity, and density distribution, these quantitative features could also serve as structured inputs for predictive models. This opens the possibility of developing statistical risk models trained on large numbers of simulated or retrospective cases, complementary to the FE-based BOS score. In addition, future developments could explore unsupervised or self-supervised learning methods capable of identifying abnormal bone patterns directly from imaging data, without requiring explicit segmentation masks [15]. Initial studies in other cancer domains support this direction: for instance, researchers have applied generative models and anomaly detection frameworks to identify metastatic lesions as deviations from normal anatomy in whole-body scans, without any manual labels [15]. AI models that learn to distinguish between healthy and pathological bone based on contextual density distributions or underlying structural features may offer a more robust pathway for fully automated, patient-specific risk estimation.

Overall, the findings of this thesis support the potential role of AI in improving the efficiency and standardization of lesion segmentation for fracture risk assessment in patients with femoral bone metastases. While the AI-assisted lesion segmentation (integrated into BOS score calculation) showed promising results, segmentation variability remains a critical factor in individual cases. Therefore, although AI-driven segmentation offers advantages at the group level, its integration into patient-specific modeling requires cautious interpretation and further validation to ensure reliability across a broad clinical population. At the same time, improving the quality of the ground truth itself will be essential. Standardized multi-reader annotations, supported by consensus procedures and the use of combined CT-MRI information during contouring, could result in more consistent reference segmentations for training and validation. In particular, improving generalizability will require training on larger and more diverse datasets, ideally collected across multiple centers to better reflect the variability of clinical presentations. However, it should also be acknowledged that currently-used clinical tools such as Mirels' score, axial cortical involvement, and BOS score where it is validated (e.g., in osteolytic lesions) are known to have limited specificity and therefore the performance of AI-assisted approaches should be evaluated relative to these existing standards rather than against an unrealistic expectation of perfect accuracy for every patient.

Improvements in FE simulations

Accurate assessment of fracture risk using FE modeling relies strongly on proper calibration of CT scan data to estimate local bone density and mechanical properties. Traditionally, this has required a calibration phantom scanned alongside the patient. However, this setup is often unavailable in retrospective datasets and can be logistically impractical in routine clinical workflows. Moreover, the absence of calibration phantoms in retrospective cohorts limits the possibility of validating fracture risk models against real-world patient outcomes.

To address this limitation, the air-fat-muscle (AFM) method was adopted and evaluated in this thesis (Chapter 5) as a phantomless calibration approach. By using internal tissue references visible on standard CT scans (air, fat, and muscle), the AFM method enables bone density calibration without the need for external phantoms [16]. This approach not only simplifies data acquisition but also enables the inclusion of retrospective imaging databases in validation studies, particularly those that include clinical follow-up data on fracture occurrence, thereby expanding the scope of clinical evaluation.

A study by Bartenschlager et al. compared the performance of four phantomless calibration methods (including the AFM method) with the gold standard of simultaneous calibration and showed that phantomless approaches pre-calibrated on a reference dataset achieved very small BMD differences from phantom-based calibration, while methods lacking such pre-calibration performed worse [17]. Damron et al. also reported that AFM calibration method has already

been applied to QCT-FE fracture risk assessment in metastatic femurs, with no detectable difference in predicted failure load versus phantom-based calibration [18]. Furthermore, it is more compatible with standard imaging protocols used in most clinical settings. The robustness of the AFM method was further examined across a range of clinical CT protocols. The approach has also been extended beyond the femur: Zhu et al. used the AFM calibration method to derive element-specific bone moduli in a patient-specific pelvic FE model, confirming its practical applicability for complex skeletal reconstructions without a calibration phantom [19].

Results of Chapter 5 demonstrated that, despite variations in scanning parameters, the AFM approach produced failure load estimates widely similar to phantom calibration for most protocol variations. However, notable deviations occurred when tube voltage was altered and in a specimen with muscle degeneration. Given the limited sample size, these results should be considered preliminary. Accordingly, AFM and phantom calibration should be regarded as comparable only when acquisition protocols are harmonized and tissue composition is typical, rather than generally interchangeable.

Importantly, patient-specific phantomless calibration removes a significant barrier to routine fracture risk screening. Prado et al. showed that using internal fat and muscle as reference points can yield bone density estimates with reliability comparable to phantom-based methods [20]. Their study demonstrated that subject-specific phantomless calibration using internal reference tissues (fat and muscle) can closely reproduce phantom-based BMD and vertebral failure load estimates. It also showed moderate correlations with DXA. In addition, Eggermont et al. developed and tested AFM calibration method for metastatic femoral FE modeling under real-world multi-scanner conditions, and highlighted the relative robustness of AFM compared with non-patient-specific approaches [16]. This advancement enables leveraging large retrospective CT cohorts for model validation against actual fracture outcomes where previous retrospective analyses were limited by missing calibration phantoms and inconsistent scan parameters.

Recent studies also suggest that AI could further facilitate this calibration step. For instance, deep learning models can now estimate volumetric bone mineral density directly from routine CT scans (achieving approximately 0.9 correlation with DXA measurements) [21]. However, the AI-based BMD estimation approach presented in [21] was developed in osteoporotic cohorts and have not yet been validated in patients with femoral metastases, where pathological bone remodeling may alter the relationship between radiodensity and mechanical properties. Additionally, a recent study applied a deep learning model to estimate vertebral volumetric BMD in patients with spinal metastatic tumors and demonstrated strong agreement with QCT-derived reference values (Spearman correlation coefficient of 0.92), indicating that AI-based opportunistic BMD estimation may also hold potential in cancer patients [22]. Another recent work showed that convolutional neural networks trained on a large FE-generated dataset of 2D

projections (36,000 synthetics with lytic lesions) can predict experimentally measured femoral strength with high accuracy ($R^2 \geq 0.80$), outperforming 2D densitometry and approaching 3D FE results [23].

FE modeling in this thesis was also adapted to better simulate the biomechanical impact of metastatic lesions, particularly osteoblastic cases, while for osteolytic lesions, the use of manual versus automatic segmentation was investigated. Purely osteolytic metastases aggressively weaken bone structure and are associated with the highest fracture rates among lesion types [24]. Therefore, segmented lesions were modeled with zero mechanical strength (Chapter 3) to represent local structural weakness. This approach ensured that the loss of structural integrity was not underestimated, particularly for lesions located in regions critical to load transfer.

In osteoblastic lesions, a different modeling challenge arises. Here, the increased radiodensity visible on CT does not necessarily indicate higher mechanical strength due to the presence of disorganized or sclerotic bone [25, 26]. To address this, lesion-specific material properties were adjusted to reduce potential overestimation of local strength (Chapter 4). This aligns with prior observations that standard CT-to-BMD mappings derived from healthy bone may not reflect the altered composition of metastatic tissue, often leading to overestimated strength in pathological bone. Historically, FE analyses simply applied healthy bone density-elasticity relationships even in tumor-affected regions [24, 27], which may results in overestimation of local strength.

While these refinements did not drastically change predicted failure loads at the group level, their influence was more pronounced in specific cases. In particular, lesions situated in high-stress regions of the bone had a larger impact on the predicted strength, and this effect is especially relevant for BOS scores that lie close to the clinical decision threshold, where small changes in input can shift a patient from low- to high-risk classification. This suggests that accurate material property assignment becomes increasingly important in such scenarios. Although we cannot claim that these adaptations fully capture the real-world mechanical behavior of bones with metastases, they improve the anatomical and biomechanical validity of the models. Further validation studies, ideally incorporating experimental fracture data or clinical outcomes, will be necessary to fully substantiate their impact. A remaining difficulty is the lack of empirical data on the mechanical properties of metastatic bone tissue, making it hard to definitively calibrate our model's material assumptions. Direct mechanical testing of metastatic tissue obtained from surgical biopsies or cadaveric specimens could provide quantitative strength-density relationships specific to pathological bone, which would enable more accurate calibration of FE material models [28]. Advanced imaging techniques, such as high-resolution peripheral quantitative CT or ultra-high-field MRI, may also offer alternative measures of microstructural integrity that correlate with mechanical competence [29, 30]. Additionally, future work might leverage AI pattern recognition to infer tissue quality from radiological features.

In summary, the adoption of the AFM calibration method and the tailored modeling of lesion types represent important steps toward making FE simulations more clinically feasible and biomechanically accurate. These developments support broader implementation of FE modeling for patient-specific fracture risk assessment in metastatic bone disease. Larger prospective studies will be needed to establish definitive risk thresholds and to gain clinical acceptance. Additionally, combining biomechanical risk assessment with patient-specific factors (such as tumor progression and overall prognosis) possibly via AI-driven decision support [31] could help reduce both undertreatment and unnecessary surgery compared to decisions based solely on imaging criteria.

Clinical relevance

The BOS score was explored in this thesis as a potential tool for patient-specific fracture risk assessment. Compared with traditional methods such as Mirels' score or axial cortical involvement, the BOS score demonstrated superior sensitivity and specificity [32], thereby reducing the likelihood of both unnecessary surgical interventions and missed opportunities for timely preventive fixation. While no method, including Mirels' score or axial cortical involvement, can predict fractures with absolute certainty, the BOS score offers a more quantitatively grounded estimate of mechanical failure risk. The fact that patients with low BOS scores may not sustain a fracture does not necessarily imply that the risk estimate was incorrect. Fracture occurrence is strongly influenced by patient-specific factors beyond bone strength, including physical activity levels, falls, or concurrent treatments that alter bone quality [33, 34]. Consequently, a low BOS score should not be interpreted as predictor of fracture, but rather as an indication that fracture is more likely under normal physiological loading conditions.

Conversely, the quantification of a high BOS score (i.e. a low fracture risk) for a particular patient can provide reassurance to both the patient and the clinician [32]. Patients who are informed that their fracture risk is low may feel more confident in maintaining physical activity or participating in exercise programs without fear of imminent fracture. This reassurance can be clinically valuable, as physical activity is often essential for preserving mobility, muscle strength, and overall quality of life in cancer patients [33]. In this way, the BOS score can potentially function not only as a predictive tool but also as a facilitator of patient empowerment and shared decision-making.

Overall, the clinical relevance of the work as described in this thesis lies not in delivering a ready-to-use tool, but in demonstrating the obstacles that must be addressed before FE-based scores can become reliable for clinical decision-making, particularly for patients with osteoblastic lesions. While lesion-specific modeling and AI-based segmentation represent potential ways forward, their current limitations, particularly the inconsistency of segmentation accuracy and

the lack of reliable data for osteoblastic lesions, mean that their clinical impact remains limited at this point.

Regulatory perspective

The BOS score calculation workflow presented in this thesis combines several automated steps, including AI-based lesion segmentation, phantomless CT scan calibration, and FE simulations. The goal is to support clinicians in assessing fracture risk in femora affected by metastatic bone disease. Importantly, this workflow is not intended to make clinical decisions independently. Instead, it provides decision-support by generating standardized, reproducible data that clinicians can interpret alongside other clinical findings. Within this workflow, only the lesion segmentation component involves learning-based outputs, as it uses trained AI models to segment metastatic lesions. The remaining components are based on established deterministic methods and serve to automate processes that have already been validated in prior studies [16, 32].

From a regulatory standpoint, this type of software would fall under the definition of Software as a Medical Device (SaMD), according to international guidance such as that from the International Medical Device Regulators Forum (IMDRF). In the United States, the Food and Drug Administration (FDA) assesses such tools using a risk-based approach that considers how the software supports clinical management and its potential to impact patient safety. In the European context, both the Medical Device Regulation (MDR) and the upcoming Artificial Intelligence (AI) Act are relevant. While the deterministic steps in the BOS workflow may face fewer regulatory barriers, the AI-based lesion segmentation module would likely receive closer attention. This is due to its dependence on machine learning and the way it affects the interpretation of fracture risk outcomes.

Regulatory requirements for such software would include validation of its clinical performance, safety, and reproducibility in real-world settings. For the AI component, this includes demonstrating reliable segmentation performance across different lesion types and anatomical regions, and across a range of CT imaging protocols. In addition, transparency and explainability are growing expectations for AI-enabled systems. Therefore, the process by which the AI model produces lesion contours should be documented in a way that is interpretable to clinicians and regulatory reviewers. Furthermore, post-market surveillance would be required to monitor how the system performs over time, especially when deployed in diverse clinical environments.

Since the BOS workflow is designed to assist clinicians and explicitly not to make autonomous decisions, its regulatory classification may be less complex than that of fully independent AI diagnostic tools. Its primary function is to provide biomechanical insight that complements, rather than replaces, clinical judgment. This role is consistent with evolving regulatory principles

that support the integration of AI into clinical workflows, provided that human oversight is maintained throughout the decision-making process.

Limitations

While this thesis advances the clinical applicability of FE modeling for fracture risk assessment in femoral metastases, several limitations remain. A key challenge lies in the accurate segmentation of bone lesions. Although, relative to the manual workflow, the use of deep learning, particularly nnU-Net-based models, substantially improved the efficiency and reproducibility of lesion delineation, segmentation of metastatic bone disease remains complex. Lesions are often irregular in shape, variable in size, and may extend into both cortical and trabecular compartments, complicating automated boundary detection. Furthermore, differentiating between osteolytic, osteoblastic, and mixed patterns on CT scans is not always straightforward, which limits the precision of automated algorithms.

The development of reliable AI-driven segmentation models is also constrained by the availability of training data. High-quality, expert-annotated datasets for bone metastases are relatively limited compared with those for soft-tissue tumors. This deficiency limits the diversity of cases that can be used to train and validate deep learning models, potentially affecting their ability to generalize to new patient populations or imaging protocols. Expanding such datasets across institutions and scanners remains an important step toward ensuring robust and widely applicable AI tools.

Another limitation is the lack of detailed data on the material properties of osteoblastic lesions. While this thesis incorporated lesion-specific adjustments to account for the mechanical weakness of these lesions, their true biomechanical behavior is not yet fully understood. Histological studies have shown that osteoblastic lesions often consist of dense but disorganized bone, which appears strong on CT but is mechanically fragile [35, 36]. However, quantitative measurements that link radiological density to actual mechanical performance are still limited. This gap constrains the accuracy of current FE simulations and underlines the need for future experimental and clinical studies to better characterize these lesions.

In summary, these limitations indicate areas where further research is needed to enhance both the accuracy and the clinical applicability of FE-based fracture risk assessment. Addressing them will be essential for translating the methods developed in this thesis into routine clinical practice.

Future work

The findings of this thesis highlight several promising directions for further research and development to enhance fracture risk assessment in patients with femoral bone metastases. A key area for future progress lies in advancing AI-based lesion segmentation. While this thesis

demonstrated the feasibility of using deep learning frameworks such as nnU-Net for femoral metastases, AI in medical imaging is a rapidly evolving field. Over the past few years, new architectures have emerged, including transformer-based models and foundation models such as the Segment Anything Model (SAM), which allow broader generalization across imaging modalities and tasks [37, 38]. These methods hold the potential to further improve segmentation accuracy, particularly for complex lesions such as diffuse or mixed-density metastases, and could reduce the need for manual correction. Continued integration of these advances into skeletal oncology will be critical to ensuring that lesion segmentation remains both accurate and efficient as AI technology matures.

Another important direction involves improving the characterization of the mechanical properties of osteoblastic lesions. Current FE models are limited by the assumption that radiodensity correlates with bone strength, which does not hold true for osteoblastic tissue. Experimental studies on donor femurs affected by osteoblastic metastases would provide the most direct insight, but such investigations are extremely challenging. Obtaining sufficiently large and representative samples is difficult, given the rarity of suitable donor material and the ethical and logistical constraints involved. Collaborative efforts at a national or European level could facilitate access to a larger cohort of patients and tissue samples, which would substantially strengthen future validation efforts. As a result, indirect approaches such as combining imaging biomarkers with histological and clinical data may provide the most realistic path forward in the near term.

Beyond refining lesion-specific FE modeling, an alternative approach would be to develop a statistical model to predict fracture risk, potentially reducing the need for time-intensive FE simulations. Unlike a purely AI-driven model, a statistical approach could be based on systematically generated data from a very large set of FE simulations. For example, one could simulate with FE modeling a million femurs with varying lesion types, sizes, and locations, and then train a predictive model to estimate fracture risk directly. Comparable methodologies have been applied in other domains of biomedical science, where simplified predictive models are trained on physics-based simulations to reduce computational requirements [39, 40]. However, the greatest challenge lies in ensuring that the training set captures the full diversity of clinically relevant lesions. Without adequate representation, the model risks poor generalizability, particularly in rare or atypical cases. If successful, such a model could be implemented as a Picture Archiving and Communication System (PACS)-integrated software tool, providing clinicians with rapid, point-of-care fracture risk assessments.

The ultimate vision is to achieve full automation of the patient-specific BOS score workflow. This would include AI-driven lesion and bone segmentation, automated calibration, optimized material property assignment, and streamlined FE simulation, all integrated into a seamless

platform. Such a system could support widespread adoption of patient-specific fracture risk assessment by minimizing operator dependency and reducing analysis time.

Finally, while the AFM calibration method already offers a practical and reliable solution for phantomless CT calibration, future developments may allow AI to further refine this process. AI could be used to learn calibration mappings directly from large multi-center datasets and compensate for protocol-related intensity shifts, effectively harmonizing BMD estimation across scanners and acquisition settings. In this scenario, the added value of AI would not lie in replacing the AFM method itself but in reducing inter-scanner variability and improving consistency when applying phantomless calibration in diverse clinical environments.

CONCLUSION

This thesis advances the clinical applicability of FE modeling for fracture risk assessment in patients with femoral bone metastases. By integrating deep learning-based lesion segmentation, adapting material properties to reflect the biomechanical characteristics of metastatic tissue, and implementing a robust phantomless CT calibration method, the research addressed key limitations that currently restrict the extension of patient-specific modeling to osteoblastic and mixed lesions in clinical practice. The results, however, showed that the approaches that were tested in this thesis are not yet sufficiently reliable for clinical use. Manual segmentation was inconsistent, AI-driven segmentation with nnU-Net showed variable accuracy, and adaptations of lesion-specific properties only modestly affected predictive performance and in some cases even led to less accurate assessments. Phantomless calibration worked appropriately under most conditions but was sensitive to certain scan and tissue parameters.

These findings suggest that the BOS score, in its current form, should be considered exploratory for patients with osteoblastic or mixed lesions. This thesis highlights both the opportunities and the remaining challenges of applying FE-based fracture risk assessment in metastatic bone disease. The studies also provide insight into where uncertainty arises in current workflows and why clinical translation is still limited. Since fracture occurrence is influenced by factors beyond bone strength, including physical activity, accidental falls, and systemic treatments, the BOS score remains an appropriate decision-support tool for patients with osteolytic lesions, where its clinical use has already been established. For patients with osteoblastic and mixed lesions, the results presented in this thesis indicate that the BOS score has the potential to serve as a valuable support alongside existing clinical criteria, but its application should currently be considered exploratory.

In summary, this thesis highlights the value of integrating advanced image analysis with biomechanical simulation to enable more personalized and evidence-based patient care for patients with bone metastases. The work described in this thesis incorporates progressive steps toward extending BOS-based fracture risk assessment beyond osteolytic lesions, by addressing its current limitations in osteoblastic and mixed lesions, particularly the challenge of achieving reliable assessments in complex (osteoblastic and mixed) lesions. Our results indicate that fracture risk assessments are generally driven more by overall bone quality than by the exact material properties assigned to osteoblastic lesions. Still, some cases did show sensitivity to these parameters, and such cases cannot be identified in advance. This highlights the need for assessment methods that are robust for every patient, not only for the majority. These insights offer a balanced and realistic view of the current state of FE-based fracture risk assessment. As survival continues to improve for many cancer patients, with registry data from the Netherlands showing that 5-year survival for de novo metastatic cancers has improved by up to approximately 45% points over the past three decades [41], maintaining quality of life through the prevention of pathological fractures becomes increasingly critical. The methodologies developed here represent a further step toward that goal, providing the foundation for future developments such as unsupervised AI for lesion segmentation, and statistical models trained on large-scale FE datasets, which may ultimately enable widespread clinical implementation of patient-specific fracture risk assessment.

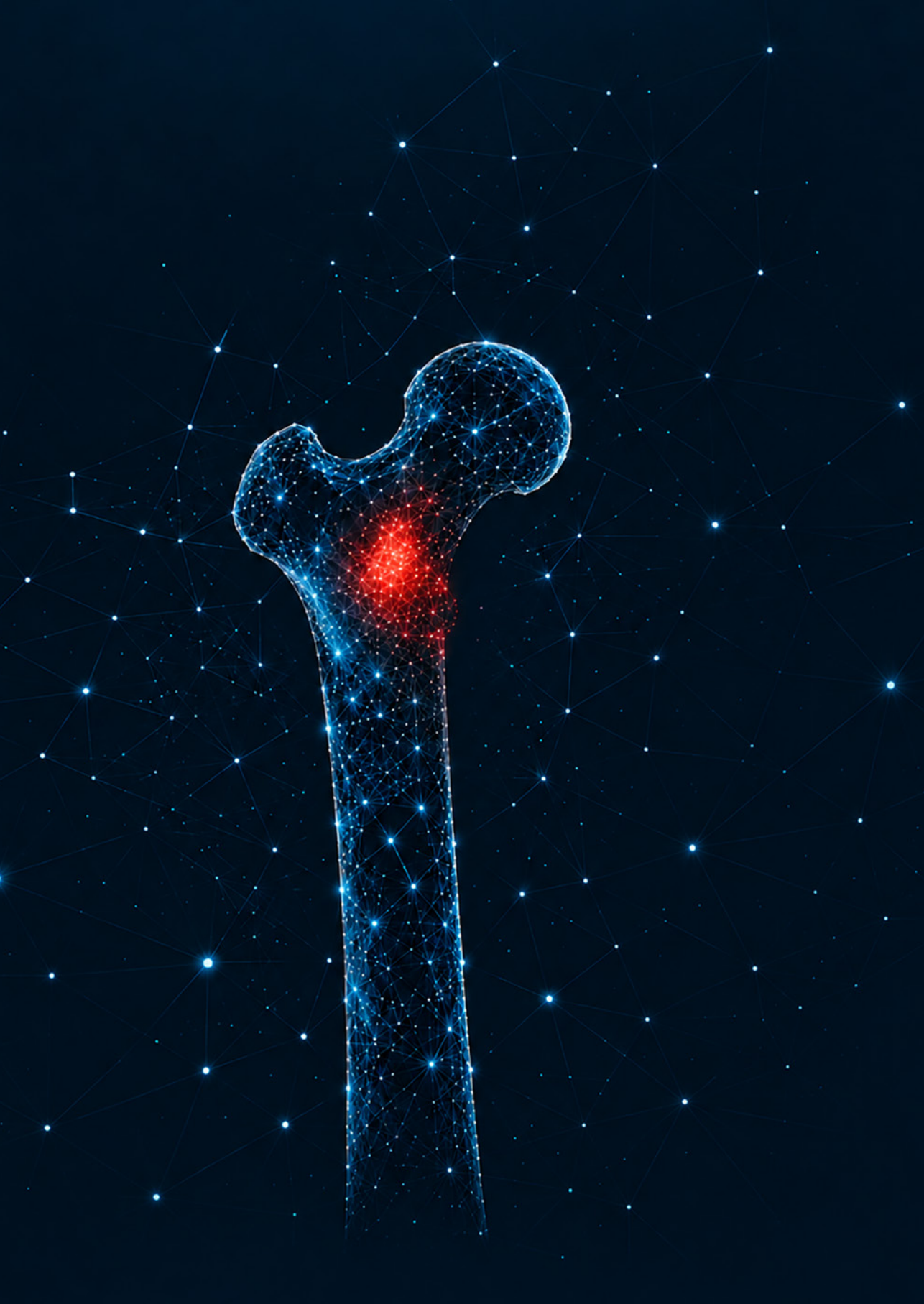
REFERENCES

1. Howard, E., K. Shepherd, G. Cribb, and P. Cool, *The validity of the Mirels score for predicting impending pathological fractures of the lower limb.* The Bone & Joint Journal, 2018. **100**(8): p. 1100-1105.
2. Van der Linden, Y., P. Dijkstra, H. Kroon, J. Lok, E. Noordijk, J. Leer, and C. Marijnen, *Comparative analysis of risk factors for pathological fracture with femoral metastases: results based on a randomised trial of radiotherapy.* The Journal of Bone & Joint Surgery British Volume, 2004. **86**(4): p. 566-573.
3. Tingelhoff, K., K.W. Eichhorn, I. Wagner, M.E. Kunkel, A.I. Moral, M.E. Rilk, F.M. Wahl, and F. Bootz, *Analysis of manual segmentation in paranasal CT images.* European archives of oto-rhino-laryngology, 2008. **265**(9): p. 1061-1070.
4. Ashton, E.A., C. Takahashi, M.J. Berg, A. Goodman, S. Totterman, and S. Ekholm, *Accuracy and reproducibility of manual and semiautomated quantification of MS lesions by MRI.* Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine, 2003. **17**(3): p. 300-308.
5. Lawrence, L.S., L.C. Chin, R.W. Chan, T.K. Nguyen, A. Sahgal, C.-L. Tseng, and A.Z. Lau, *Method of computing direction-dependent margins for the development of consensus contouring guidelines.* Radiation Oncology, 2021. **16**(1): p. 71.
6. Shen, D., G. Wu, and H.-I. Suk, *Deep learning in medical image analysis.* Annual review of biomedical engineering, 2017. **19**(1): p. 221-248.
7. Isensee, F., P.F. Jaeger, S.A. Kohl, J. Petersen, and K.H. Maier-Hein, *nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation.* Nature methods, 2021. **18**(2): p. 203-211.
8. Saha, A., L.J. Grimm, M. Harowicz, S.V. Ghate, C. Kim, R. Walsh, and M.A. Mazurowski, *Interobserver variability in identification of breast tumors in MRI and its implications for prognostic biomarkers and radiogenomics.* Medical physics, 2016. **43**(8Part1): p. 4558-4564.
9. Gerlich, A.S., J.M. van der Velden, A.N. Kotte, C.-L. Tseng, G. Fanetti, W.S. Eppinga, N. Kasperts, M.P. Intven, F.A. Pameijer, and M.E. Philippens, *Inter-observer agreement in GTV delineation of bone metastases on CT and impact of MR imaging: A multicenter study.* Radiotherapy and Oncology, 2018. **126**(3): p. 534-540.
10. Nichyporuk, B., J. Cardinell, J. Szeto, R. Mehta, J.-P.R. Falet, D.L. Arnold, S.A. Tsaftaris, and T. Arbel, *Rethinking generalization: The impact of annotation style on medical image segmentation.* arXiv preprint arXiv:2210.17398, 2022.
11. Rachmil, O., M. Artzi, M. Iluz, I. Druckmann, Z. Yosibash, and A. Sternheim, *Automatic segmentation of femoral tumors by nnU-net.* Clinical Biomechanics, 2024. **116**: p. 106265.
12. Rachmil, O., K. Myers, O. Merose, A. Sternheim, and Z. Yosibash, *The influence of femoral lytic tumors segmentation on autonomous finite element analysis.* Clinical Biomechanics, 2024. **112**: p. 106192.
13. Salembier, C., G. Villeirs, B. De Bari, P. Hoskin, B.R. Pieters, M. Van Vulpen, V. Khoo, A. Henry, A. Bossi, and G. De Meerleer, *ESTRO ACROP consensus guideline on CT-and MRI-based target volume delineation for primary radiation therapy of localized prostate cancer.* Radiotherapy and Oncology, 2018. **127**(1): p. 49-61.
14. Kohl, S., B. Romera-Paredes, C. Meyer, J. De Fauw, J.R. Ledsam, K. Maier-Hein, S. Eslami, D. Jimenez Rezende, and O. Ronneberger, *A probabilistic u-net for segmentation of ambiguous images.* Advances in neural information processing systems, 2018. **31**.

15. Patel, A., P.-D. Tudosiu, W.H.L. Pinaya, G. Cook, V. Goh, S. Ourselin, and M.J. Cardoso. *Cross attention transformers for multi-modal unsupervised whole-body pet anomaly detection*. in *MICCAI Workshop on Deep Generative Models*. 2022. Springer.
16. Eggermont, F., N. Verdonschot, Y. Van Der Linden, and E. Tanck, *Calibration with or without phantom for fracture risk prediction in cancer patients with femoral bone metastases using CT-based finite element models.* PLoS One, 2019. **14**(7): p. e0220564.
17. Bartenschlager, S., A. Cavallaro, T. Pogarell, O. Chaudry, M. Uder, S. Khosla, G. Schett, and K. Engelke, *Opportunistic screening with CT: comparison of phantomless BMD calibration methods.* Journal of Bone and Mineral Research, 2023. **38**(11): p. 1689-1699.
18. Damron, T.A. and K.A. Mann, *Fracture risk assessment and clinical decision making for patients with metastatic bone disease.* Journal of Orthopaedic Research®, 2020. **38**(6): p. 1175-1190.
19. Zhu, Y., A. Babazadeh-Naseri, M.R. Brake, J.E. Akin, G. Li, V.O. Lewis, and B.J. Fregly, *Evaluation of finite element modeling methods for predicting compression screw failure in a custom pelvic implant.* Frontiers in Bioengineering and Biotechnology, 2024. **12**: p. 1420870.
20. Prado, M., S. Khosla, C. Chaput, and H. Giambini, *Opportunistic application of phantom-less calibration methods for fracture risk prediction using QCT/FEA.* European radiology, 2021. **31**(12): p. 9428-9435.
21. Kang, J.-W., C. Park, D.-E. Lee, J.-H. Yoo, and M. Kim, *Prediction of bone mineral density in CT using deep learning with explainability.* Frontiers in Physiology, 2023. **13**: p. 1061911.
22. Wang, Z., Y. Tan, K. Zeng, H. Tan, P. Xiao, and G. Su, *Bone density measurement in patients with spinal metastatic tumors using chest quantitative CT deep learning model.* Journal of Bone Oncology, 2024. **49**: p. 100641.
23. Synek, A., E. Benca, R. Licandro, L. Hirtler, and D.H. Pahr, *Predicting strength of femora with metastatic lesions from single 2D radiographic projections using convolutional neural networks.* Computer Methods and Programs in Biomedicine, 2025. **265**: p. 108724.
24. Sas, A., E. Tanck, A. Sermon, and G.H. van Lenthe, *Finite element models for fracture prevention in patients with metastatic bone disease. A literature review.* Bone Reports, 2020. **12**: p. 100286.
25. Chen, X., J.C.H. Goh, S.H. Teoh, S.D. De, R. Soong, and T. Lee, *Localized sclerotic bone response demonstrated reduced nanomechanical creep properties.* journal of the mechanical behavior of biomedical materials, 2013. **17**: p. 198-208.
26. Macedo, F., K. Ladeira, F. Pinho, N. Saraiva, N. Bonito, L. Pinto, and F. Gonçalves, *Bone metastases: an overview.* Oncology reviews, 2017. **11**(1).
27. Greve, T., N.M. Rayudu, M. Dieckmeyer, C. Boehm, S. Ruschke, E. Burian, C. Kloth, J.S. Kirschke, D.C. Karampinos, and T. Baum, *Finite Element Analysis of Osteoporotic and Osteoblastic Vertebrae and Its Association With the Proton Density Fat Fraction From Chemical Shift Encoding-Based Water-Fat MRI–A Preliminary Study.* Frontiers in Endocrinology, 2022. **13**: p. 900356.
28. Bailey, S., M.A. Stadelmann, P.K. Zysset, D. Vashishth, and R.N. Alkalay, *Influence of metastatic bone lesion type and tumor origin on human vertebral bone architecture, matrix quality, and mechanical properties.* Journal of Bone and Mineral Research, 2020. **37**(5): p. 896-907.
29. Zhou, B., J. Wang, Y.E. Yu, Z. Zhang, S. Nawathe, K.K. Nishiyama, F.R. Rosete, T.M. Keaveny, E. Shane, and X.E. Guo, *High-resolution peripheral quantitative computed tomography (HR-pQCT) can assess microstructural and*

biomechanical properties of both human distal radius and tibia: Ex vivo computational and experimental validations. Bone, 2016. **86**: p. 58-67.

30. Jerban, S., S. Alenezi, A.M. Afsahi, Y. Ma, J. Du, C.B. Chung, and E.Y. Chang, *MRI-based mechanical competence assessment of bone using micro finite element analysis (micro-FEA).* Magnetic resonance imaging, 2022. **88**: p. 9-19.
31. Nguyễn, M.-V., C. Carlier, C. Nich, F. Gouin, and V. Crenn, *Fracture risk of long bone metastases: a review of current and new decision-making tools for prophylactic surgery.* Cancers, 2021. **13**(15): p. 3662.
32. Eggermont, F., Y. van der Linden, N. Verdonschot, E. Dierselhuis, S. Ligthert, T. Bitter, P. Westhoff, and E. Tanck, *A Patient-Specific Fracture Risk Assessment Tool for Femoral Bone Metastases: Using the Bone Strength (BOS) Score in Clinical Practice.* Cancers, 2022. **14**(23): p. 5904.
33. Rendeiro, J.A., C.A.M.P. Rodrigues, L. de Barros Rocha, R.S.B. Rocha, M.L. da Silva, and K. da Costa Cunha, *Physical exercise and quality of life in patients with prostate cancer: systematic review and meta-analysis.* Supportive care in cancer, 2021. **29**(9): p. 4911-4919.
34. Diana, A., F. Carlino, E.F. Giunta, E. Franzese, L.P. Guerrera, V. Di Lauro, F. Ciardiello, B. Daniele, and M. Orditura, *Cancer treatment–induced bone loss (CTIBL): State of the art and proper Management in Breast Cancer Patients on endocrine therapy.* Current Treatment Options in Oncology, 2021. **22**(5): p. 45.
35. Kolb, A.D. and K.M. Bussard, *The bone extracellular matrix as an ideal milieu for cancer cell metastases.* Cancers, 2019. **11**(7): p. 1020.
36. Coleman, R., *Metastatic bone disease: clinical features, pathophysiology and treatment strategies.* Cancer treatment reviews, 2001. **27**(3): p. 165-176.
37. Shamshad, F., S. Khan, S.W. Zamir, M.H. Khan, M. Hayat, F.S. Khan, and H. Fu, *Transformers in medical imaging: A survey.* Medical image analysis, 2023. **88**: p. 102802.
38. Mazurowski, M.A., H. Dong, H. Gu, J. Yang, N. Konz, and Y. Zhang, *Segment anything model for medical image analysis: an experimental study.* Medical Image Analysis, 2023. **89**: p. 102918.
39. Willard, J., X. Jia, S. Xu, M. Steinbach, and V. Kumar, *Integrating physics-based modeling with machine learning: A survey.* arXiv preprint arXiv:2003.04919, 2020. **1**(1): p. 1-34.
40. Pegolotti, L., M.R. Pfaller, N.L. Rubio, K. Ding, R.B. Brufau, E. Darve, and A.L. Marsden, *Learning reduced-order models for cardiovascular simulations with graph neural networks.* Computers in Biology and Medicine, 2024. **168**: p. 107676.
41. Luyendijk, M., O. Visser, H.M. Blommestein, I.H. de Hingh, F.J. Hoebers, A. Jager, G.S. Sonke, E.G. de Vries, C.A. Uyl-de Groot, and S. Siesling, *Changes in survival in de novo metastatic cancer in an era of new medicines.* JNCI: Journal of the National Cancer Institute, 2023. **115**(6): p. 628-635.



Chapter 7

Summary

SUMMARY

This thesis focuses on advancing the assessment of fracture risk in femoral bone metastases by combining finite element (FE) modeling with a lesion-specific material model and deep learning-based segmentation. In this chapter, an integrated overview is provided of how Chapters 2 through 5 address the main research questions, thereby illustrating the methodological developments and clinical implications of the work.

Manual Segmentation Reliability of Femoral Bone Metastases: Chapter 2

Research Question 1: What is the inter- and intra-operator reliability of manual segmentation of femoral metastatic lesions?

The first step in this thesis was to investigate the reliability of manual segmentation, which remains the reference standard for identifying metastatic lesions. Accurate segmentation of metastatic lesions is important for radiotherapy planning and for the improvement of biomechanical models that calculate the fracture risk of metastatic bones, and therefore, the study set out to assess the inter- and intra-operator reliability of this process. To do so, CT scans of 54 metastatic femurs (19 osteolytic, 17 osteoblastic, and 18 mixed) were segmented twice by two operators, and Dice coefficients (DCs) were calculated under the assumption that a DC > 0.7 indicates good reliability.

Generally, rather poor inter- and intra-operator reliability of lesion segmentation were found. Inter-operator DCs were 0.54 (±0.28) and 0.50 (±0.32) for the first and second segmentation, respectively, whereas intra-operator DCs were 0.56 (±0.28) for operator I and 0.71 (±0.23) for operator II. Larger lesions scored significantly higher DCs in comparison with smaller lesions. Of the femurs with larger mean segmentation volumes, 83% and 93% were segmented with good inter- and intra-operator DCs (>0.7), respectively. There was no difference between the mean DCs of osteolytic, osteoblastic, and mixed lesions.

In summary, these findings demonstrated that manual segmentation of femoral metastatic lesions remains highly challenging and inconsistent, highlighting the clear need for computer-assisted segmentation tools to facilitate radiotherapy planning and to improve the estimation of fracture risk.

Osteolytic lesion-specific segmentation and finite element analysis: Chapter 3

Research Question 2: How can osteolytic lesion-specific material properties be integrated into FE models?

Building upon the challenges of manual segmentation, the next step was to investigate how automatically segmented lesions could be incorporated into FE models for osteolytic and mixed metastases. In this study, we assessed the difference 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 how these differences potentially affect clinical decision-making.

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 bone mineral density (BMD) of both femur and metastatic lesion was extracted from CT-values. A self-configuring variant of the popular deep learning model U-Net, known as nnU-Net, was used to automatically identify metastatic lesions within the femur. For these models with lesions segmented by the neural network (Finite Element with Segmented Lesion: FE-with-SL), the BMD within the metastatic lesions was set to zero. The models (either with or without automatically segmented lesions) were loaded incrementally in axial direction until failure of the femur. Failure loads and patterns were determined. Furthermore, Bone Strength (BOS) scores were calculated for both the FE-no-SL and FE-with-SL, and sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were determined for both groups by comparing the predictions to the occurrence or absence of actual fracture within the patient cohort.

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 in comparison to the other cases, as for these cases, the models with automatic lesion segmentation (FE-with-SL) showed considerably lower failure loads compared to the FE-no-SL. The correlation between failure load difference and lesion size showed that, generally, larger segmented lesions led to larger differences of failure load between the two models. In the whole database including osteolytic and mixed lesions, sensitivity and NPV of the BOS score 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.

Our study showed that the nnU-Net yielded non-robust outcomes in femoral lesion segmentation, indicating that automatic lesion segmentation remains a challenging step for reliably integrating deep learning into patient-specific FE modeling. The difference in failure pattern and failure load between FE models with automatically segmented osteolytic and mixed lesions were relatively small, although a few cases showed larger deviations that could be clinically relevant.

Osteoblastic lesion-specific segmentation and finite element analysis: Chapter 4

Research Question 3: How can realistic material properties of osteoblastic lesions be integrated into FE models, and how do manual and automatic segmentation affect BOS-based fracture risk assessment?

After addressing osteolytic and mixed lesions, the research turned to the unique challenges posed by osteoblastic metastases, where increased bone density complicates reliable fracture risk assessment. FE models which are generated from CT scans can be used to predict fracture risk in metastatic femora by calculating the BOS score. However, with osteoblastic lesions, reliable assessment is challenging due to the high bone density in the lesions, thereby overestimating bone strength. Hence, osteoblastic properties should be adapted and for this, lesions need to be separately identified. However, manual segmentation of these lesions is a time-consuming and subjective task, while automatic segmentation using deep learning algorithms requires validation. This study aimed to assess how modifying the BMD values of osteoblastic lesions affects the BOS score, evaluate the reliability of automatic segmentation, and compare BOS scores derived from manual and automatic segmentation FE modeling.

Nineteen femora with osteoblastic metastases were analyzed using patient-specific FE models. The BMD of the manually segmented metastases was adjusted in steps, resulting in an incremental reduction of simulated strength from 100% to 0%. The BOS score was calculated for each step to assess its sensitivity to BMD distribution changes. Additionally, automatic segmentations were performed, and the agreement between manual and automatic segmentation was evaluated using DCs. Additionally, Bland-Altman and correlation analyses were conducted to compare BOS scores resulting from the two segmentation methods.

The BOS score generally decreased as the BMD of osteoblastic metastases was reduced. However, the relationship was not always gradual, with some femora showing minimal changes when BMD was decreased. DCs for automatic segmentation varied widely (0 to 0.88), with lower accuracy in lesions located in the femoral trochanter and diaphysis level. BOS scores based on manual and automatic segmentation were strongly correlated, and in most cases the differences were relatively small. Six femora had higher BOS scores for manual segmentation, eight for automatic, and five showed virtually no difference. In most cases, the differences were small and

not likely to be clinically relevant. However, for specific cases considerable effects were observed which may be of clinical relevance.

In conclusion, adjusting the BMD distribution of osteoblastic lesions influences BOS scores, but the effect varies based on lesion characteristics such as size and location. Automatic segmentation showed promise but exhibited variability depending on lesion location. Future research should focus on refining BMD adaptation for improved fracture risk assessment and expanding the training dataset for automatic segmentation to enhance its reliability.

Influence of CT Scan Protocol Variations: Chapter 5

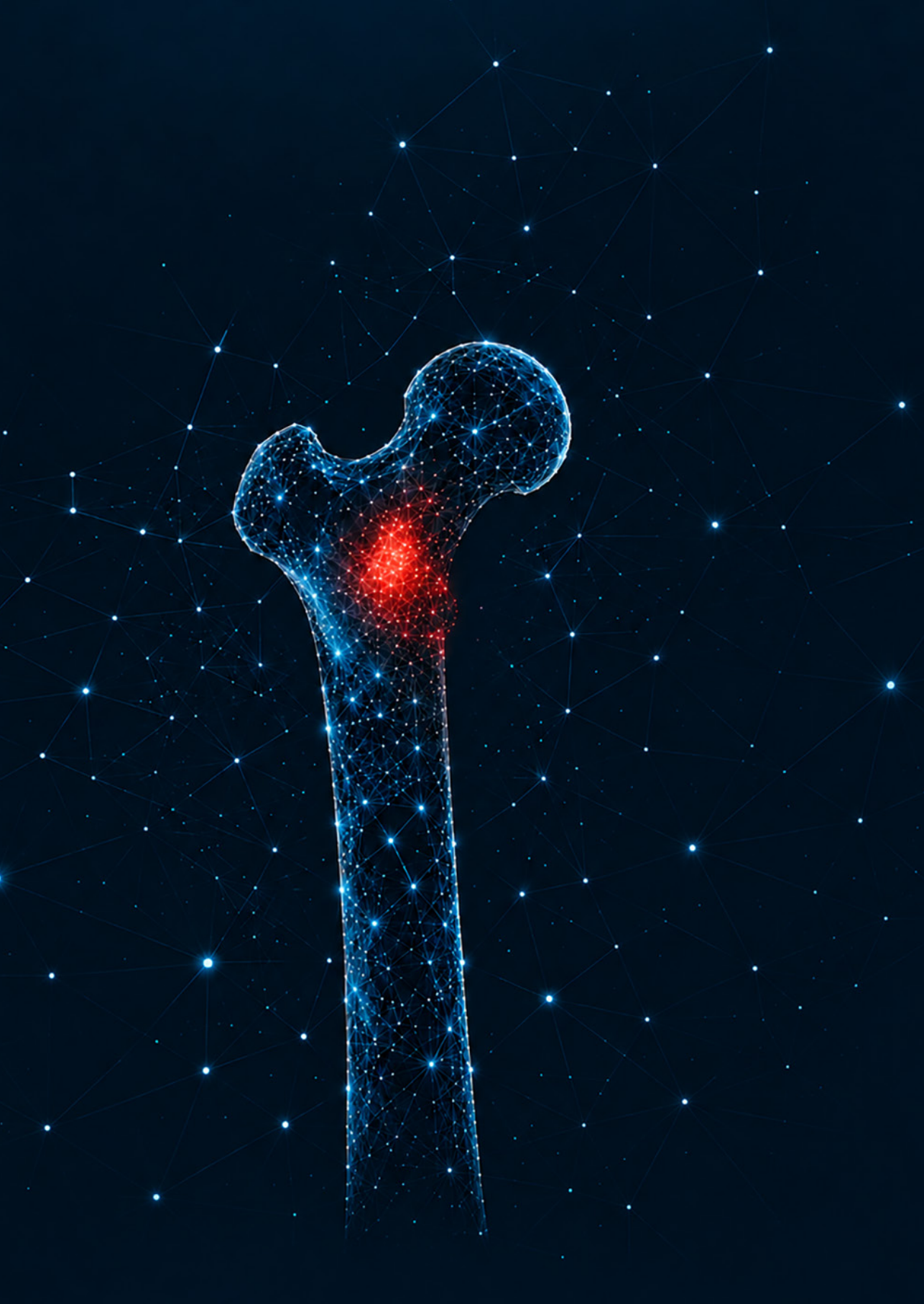
Research Question 4: How do variations in CT scan protocol affect femoral finite element failure load assessment when using air-fat-muscle calibration method?

The robustness of the FE modeling pipeline was examined by testing whether variations in CT protocols would alter the accuracy of fracture risk assessment. Recently, it was shown that fracture risk assessment using FE modeling can be performed using a calibration phantom, air-fat-muscle (AFM) calibration, or a non-patient-specific calibration approach. The performance of AFM calibration when using different CT scan protocols is, however, not known. The purpose of this study was therefore to investigate the influence of variations in CT scan protocols on Hounsfield unit (HU), BMD, and failure load (FL) between phantom, AFM calibration, and non-patient-specific calibration.

Five human cadaver lower limbs were scanned atop a calibration phantom according to a standard scanning protocol and seven additional deviating protocols including variations in current-time, tube voltage, slice thickness, rotation time, field of view, reconstruction kernel, and reconstruction algorithm. The HUs of the scans were calibrated to BMD using the calibration phantom, by AFM calibration and by non-patient-specific calibration. FE models were created, and failure loads were calculated by simulating an axial load on the femur. HU, BMD, and FL of all protocols were compared between the three calibration methods.

The different protocols showed little variation in HU, BMD and FL. Only changing the tube voltage resulted in a relatively large decrease of approximately 10% in mean HU and BMD of the trabecular and cortical region of interest (ROI), resulting in a 13.8% lower FL for AFM and 13.4% lower FL for non-patient-specific calibration. BMDs were generally 3.3% and 3.9% higher after AFM calibration and non-patient-specific calibration compared to phantom calibration. One femur showed muscle degeneration which resulted in a larger FL when using the AFM calibration compared to the phantom calibration. While strong correlations were observed between phantom and AFM calibration, these methods should be considered comparable only when similar protocols are used. However, given the limited sample size, this conclusion should

be interpreted with caution, and further validation is needed to confirm interchangeability. Protocol deviations, particularly tube voltage should be avoided as much as possible.



Chapter 8

Samenvatting

SAMENVATTING

Dit proefschrift richt zich op het verbeteren van de beoordeling van fractuurrisico bij femorale botmetastasen door het combineren van eindige elementen (EE) modellering met een laesiespecifiek materiaalmodel en deep-learninggebaseerde segmentatie. In deze samenvatting wordt een geïntegreerd overzicht gegeven van hoe de onderzoeksvragen en hoe deze in Hoofdstukken 2 tot en met 5 zijn beantwoord.

Betrouwbaarheid van handmatige segmentatie van femorale botmetastasen: Hoofdstuk 2

Onderzoeksvraag 1: Wat is de inter- en intra-operator betrouwbaarheid van handmatige segmentatie van femorale metastatische laesies?

De eerste stap in dit proefschrift was het onderzoeken van de betrouwbaarheid van handmatige segmentatie, die nog steeds de referentiestandaard vormt voor het identificeren van metastatische laesies. Nauwkeurige segmentatie is belangrijk voor radiotherapiebestralingsplanning en voor de verbetering van biomechanische modellen die het fractuurrisico van gemetastaseerd bot berekenen. Daarom werd de inter- en intra-operator betrouwbaarheid van dit proces beoordeeld. CT-scans van 54 metastatische femurs (19 osteolytisch, 17 osteoblastisch en 18 gemengd) werden tweemaal gesegmenteerd door twee operators, en Dice-coëfficiënten (DC's) werden berekend onder de aanname dat DC > 0,7 wijst op goede betrouwbaarheid.

Over het algemeen werd een matige inter- en intra-operator betrouwbaarheid vastgesteld. De inter-operator DC's waren 0,54 (±0,28) en 0,50 (±0,32) voor respectievelijk de eerste en tweede segmentatie, en de intra-operator DC's waren 0,56 (±0,28) voor operator I en 0,71 (±0,23) voor operator II. Grotere laesies resulteerden in significant hogere DC's dan kleinere laesies. Van de femurs met grotere segmentatievolumes werd 83% en 93% gesegmenteerd met goede inter- respectievelijk intra-operator DC's (>0,7). Er werd geen verschil gevonden tussen de gemiddelde DC's van osteolytische, osteoblastische en gemengde laesies.

Samengevat toont deze studie aan dat handmatige segmentatie van femorale metastatische laesies uitdagend en inconsistent is, wat de duidelijke noodzaak onderstreept van computerondersteunde segmentatietools voor radiotherapiestrategieën en voor het verbeteren van fractuurrisicoschatting.

Osteolytische laesiespecifieke segmentatie en eindige elementen analyse: Hoofdstuk 3

Onderzoeksvraag 2: Hoe kunnen osteolytische laesiespecifieke materiaaleigenschappen worden geïntegreerd in EE-modellen?

Voortbouwend op de uitdagingen van handmatige segmentatie werd vervolgens onderzocht hoe automatisch gesegmenteerde laesies kunnen worden geïntegreerd in EE-modellen voor osteolytische en gemengde metastasen. Hierbij werden EE-modellen met automatisch gesegmenteerde laesies vergeleken met modellen waarin CT-gebaseerde laesiewaarden (de gouden standaard) werden gebruikt, en werd onderzocht hoe eventuele verschillen de klinische besluitvorming kunnen beïnvloeden.

In totaal werden 50 femurs met osteolytische en gemengde metastasen opgenomen uit twee patiëntcohorten. De femurs werden gesegmenteerd uit CT-beelden en omgezet naar EE-meshes. Het materiaalgdrag werd geïmplementeerd als niet-lineair isotroop. Deze modellen zonder gesegmenteerde laesie (EE-no-SL) gebruikten lokale botmineraaldichtheid (BMD) afkomstig uit CT-waarden. Een zelfconfigurerende variant van het bekende deep learning-model U-Net, de nnU-Net, werd gebruikt om metastasen automatisch te identificeren. In modellen met gesegmenteerde laesies (EE-with-SL) werd de BMD van de laesies op nul gezet. De modellen werden vervolgens axiaal belast tot falen, en falenbelastingen en -patronen werden bepaald. BOS-scores werden berekend voor beide EE-methoden en vergeleken met klinische fractuuruitkomsten om sensitiviteit, specificiteit, positief voorspellende waarde (PPV) en negatief voorspellende waarde (NPV) te bepalen.

De verschillen in falenbelasting tussen EE-no-SL en EE-with-SL varieerden van 0% tot 48% (gemiddeld 6,6%). De correlatieanalyse liet een sterke relatie zien ($R^2 > 0,9$). Vier casussen weken duidelijk af, waarbij EE-with-SL aanzienlijk lagere falenbelastingen voorspelde. Een grotere laesiegrootte leidde doorgaans tot grotere verschillen tussen de modellen. In de volledige dataset bleven sensitiviteit en NPV gelijk, maar specificiteit en PPV daalden van 94% naar 83% en van 78% naar 54% bij gebruik van automatische segmentatie.

Deze studie toont aan dat nnU-Net niet-robuste uitkomsten oplevert bij femorale laesiesegmentatie, waardoor automatische segmentatie een kritische uitdaging blijft voor betrouwbare patiëntspecifieke EE-modellering. De verschillen in falenpatroon en -belasting waren meestal klein, maar enkele gevallen lieten potentieel klinisch relevante afwijkingen zien.

Osteoblastische laesiespecifieke segmentatie en eindige elementen analyse: Hoofdstuk 4

Onderzoeksvraag 3: Hoe kunnen realistische materiaaleigenschappen van osteoblastische laesies worden geïntegreerd in EE-modellen, en hoe beïnvloeden handmatige en automatische segmentatie de BOS-gebaseerde fractuurrisicobeoordeling?

Na de analyse van osteolytische en gemengde laesies richtte het onderzoek zich op osteoblastische metastasen, waarbij verhoogde botdichtheid de beoordeling van fractuurrisico bemoeilijkt. EE-modellen gebaseerd op CT-scans kunnen het fractuurrisico voorspellen via de BOS-score, maar door de hoge dichtheid binnen osteoblastische laesies wordt botsterkte vaak overschat. Daarom moeten de eigenschappen van deze laesies worden aangepast, waarvoor ze eerst moeten worden geïdentificeerd. Handmatige segmentatie is tijdrovend en subjectief, terwijl automatische deep-learningsegmentatie validatie vereist.

In deze studie werden 19 femurs met osteoblastische metastasen geanalyseerd met patiëntspecifieke EE-modellen. De BMD van handmatig gesegmenteerde metastasen werd stapsgewijs aangepast, van 100% tot 0%, om de gevoeligheid van de BOS-score voor variaties in BMD te evalueren. Automatische segmentaties werden uitgevoerd en de overeenstemming met handmatige segmentatie werd beoordeeld met DC's. Bland–Altman- en correlatieanalyses werden gebruikt om BOS-scores van beide methoden te vergelijken.

De BOS-score nam over het algemeen af naarmate de BMD werd verlaagd, maar niet altijd geleidelijk. DC's varieerden sterk (0 tot 0,88), met lagere accuratesse in de trochanterregio en diafyse. BOS-scores uit handmatige en automatische segmentatie correleerden sterk en verschillen waren meestal klein en klinisch niet relevant, hoewel enkele gevallen duidelijke afwijkingen lieten zien die mogelijk klinisch betekenisvol zijn.

Conclusie: het aanpassen van de BMD binnen osteoblastische laesies beïnvloedt de BOS-score, maar het effect hangt af van laesiegrootte en -locatie. Automatische segmentatie toont potentie maar vertoont variabiliteit afhankelijk van de laesielocatie. Toekomstig onderzoek moet zich richten op verdere verfijning van BMD-aanpassing en uitbreiding van trainingsdata voor automatische segmentatie.

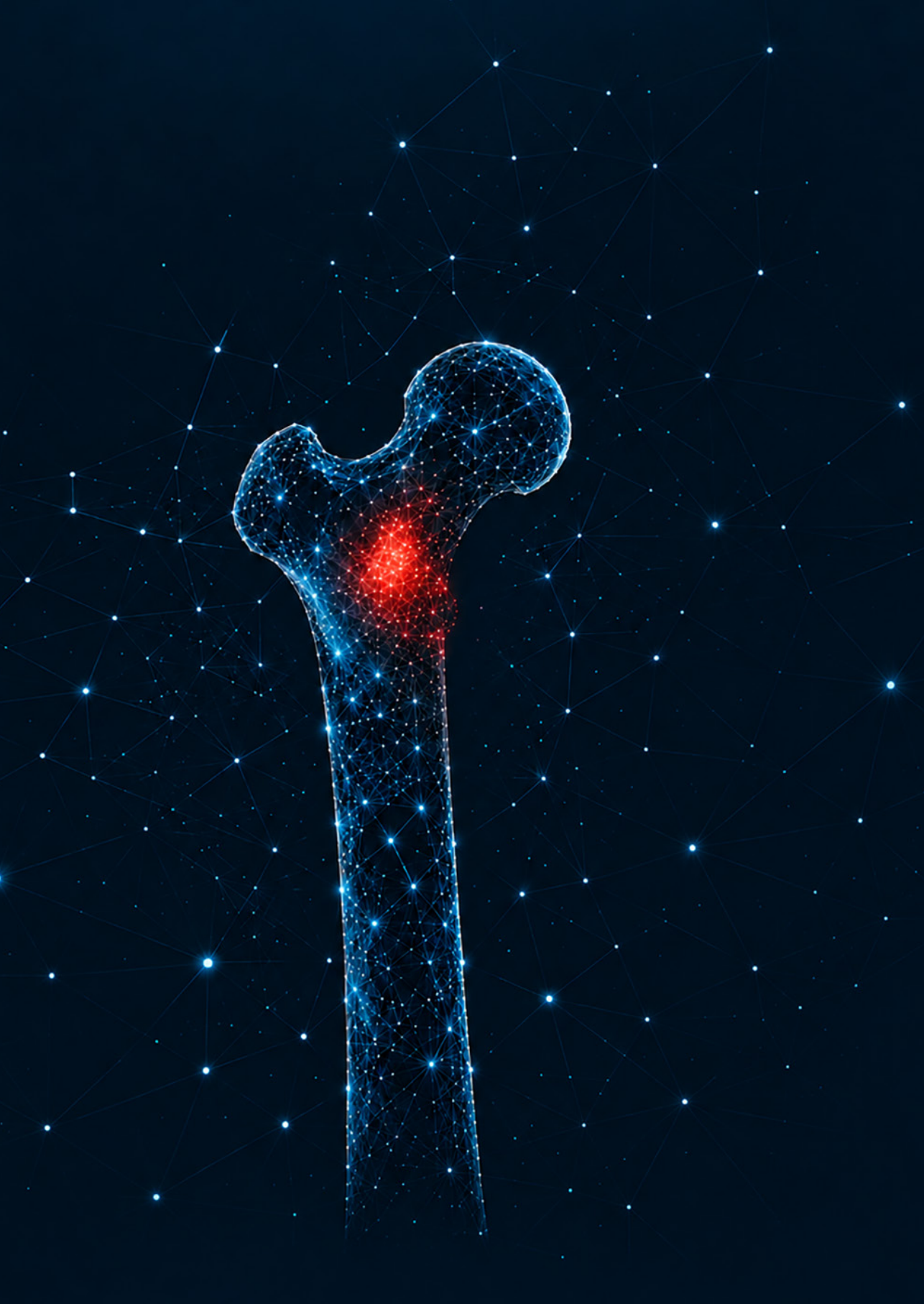
Invloed van variaties in CT-scanprotocol: Hoofdstuk 5

Onderzoeksvraag 4: Hoe beïnvloeden variaties in CT-scanprotocol de EE-berekening van falenbelasting wanneer de air–fat–muscle (AFM) kalibratiemethode wordt gebruikt?

De robuustheid van de EE-pijplijn werd onderzocht door te bepalen of variaties in CT-protocollen de nauwkeurigheid van fractuurrisicobeoordelingen beïnvloeden. Recent onderzoek heeft aangetoond dat EE-gebaseerde fractuurrisicoanalyse kan worden uitgevoerd met een kalibratiefantoom, AFM-kalibratie of een niet-patiëntspecifieke kalibratie. De prestaties van AFM-kalibratie bij verschillende CT-instellingen waren echter niet bekend.

Vijf menselijke cadaverbenen werden gescand op een kalibratiefantom volgens een standaardprotocol en zeven afwijkende protocollen (variaties in stroomsterkte, buisspanning, snijdikte, rotatietijd, field of view, reconstructiekernel en reconstructiealgoritme). Hounsfield Units (HU) werden omgezet naar BMD via het fantoom, AFM-kalibratie en niet-patiëntspecifieke kalibratie. EE-modellen werden gegenereerd en falenbelastingen berekend via axiale belasting. HU, BMD en FL (failure load) werden tussen protocollen en kalibratiemethoden vergeleken.

De meeste protocolvariaties veroorzaakten slechts kleine verschillen in HU, BMD en FL. Alleen variaties in buisspanning veroorzaakten een forse daling van ~10% in HU en BMD binnen trabeculaire en corticale regio's, wat resulteerde in een 13,8% lagere FL bij AFM-kalibratie en 13,4% bij niet-patiëntspecifieke kalibratie. BMD's waren gemiddeld 3,3% en 3,9% hoger bij AFM- respectievelijk niet-patiëntspecifieke kalibratie ten opzichte van fantoomkalibratie. Eén femur met een hoge mate van spierdegeneratie liet een hogere FL zien bij AFM-kalibratie dan bij fantoomkalibratie. Hoewel sterke correlaties werden gevonden, moeten AFM- en fantoomkalibratie alleen gelijkwaardig worden beschouwd bij gelijke scanprotocollen. Gezien de beperkte steekproefomvang is voorzichtigheid geboden en is verdere validatie noodzakelijk. Vooral afwijkingen in buisspanning moeten zoveel mogelijk worden vermeden.



Appendices

Research Data Management
Radboudumc PhD Portfolio
Curriculum Vitae
Acknowledgements

RESEARCH DATA MANAGEMENT

Ethics and Privacy

The studies in this thesis used pseudonymized clinical CT scans of patients with femoral metastatic lesions. Use of these imaging datasets was permitted under existing ethical approvals from the Regionale Toetsingscommissie Patiëntgebonden Onderzoek, Leeuwarden (NL12568.099.06) and Commissie Mensgebonden Onderzoek regio Arnhem-Nijmegen (reference 2013/305). The research was conducted in accordance with applicable national and international legislation, relevant guidelines, and Radboudumc policy. Informed consent was obtained from participants to collect and process their data for this research project. Consent was not obtained for sharing and reuse of the data for future research. The privacy of participants was warranted by the use of Pseudonymized data. The research involved only secondary analysis of imaging data and computational modelling (manual and automated lesion segmentation, calibration procedures, and finite element simulations), without any additional patient interaction.

Data Collection and Storage

The data used in this thesis were clinical imaging data made for study purposes. CT scans and derived datasets such as segmentation masks, calibrated density fields, finite element models, and model outputs were processed and analyzed within the secure digital infrastructure of Radboudumc. During the research, all data were stored on secure Radboudumc department servers and managed within the Orthopaedic Research Lab (ORL), and are owned by Radboudumc. These storage facilities comply with Radboudumc data-protection requirements and ensure availability, integrity, and confidentiality through controlled access, encryption, and routine backups. Only authorized members of the research team had access to the data, ensuring confidentiality and integrity throughout the research workflow.

Data Sharing

The scientific results presented in this thesis have been communicated partly through peer-reviewed publications (Chapters 2, 3, and 5) and the remaining chapters are prepared in a format suitable for publication. The studies presented in Chapters 2, 3, and 5 have been published open access. Chapter 4 has not been published. Since the underlying CT imaging data and derived segmentations originate from clinical cohorts and are subject to medical confidentiality and data-protection regulations, they are archived with closed access for 10 years in a DAC (https://doi.org/10.34973/2g2b-rk55) of the Radboud Data Repository after termination of the study.

PHD PORTFOLIO OF ALI ATAEI

Department: Orthopaedic Research Laboratory
PhD period: 02/01/2018 – 23/11/2026

PhD Supervisor(s):
Prof. dr. ir. N.J.J. Verdonschot
Prof. dr. E.J.M. Tanck

PhD Co-supervisor(s):
Dr. N. Lessmann
Dr. F.E. Eggermont

Training activities	Hours
Courses	
RIHS - Introduction course for PhD candidates (2018)	15.00
Dutch course, Social Dutch I (2018)	36.00
Deep-learning (Eindhoven University) (2018)	16.00
Poster presentation (2019)	3.00
Improving Deep Neural Networks: Hyperparameter tuning, Regularization and Optimization (2019)	20.00
Neural Networks and Deep Learning (2019)	30.00
RU - Scientific Writing for PhD candidates (2019)	84.00
Dutch course A0-A1 (2020)	42.00
Dutch course A1-A2 (2020)	42.00
Radboudumc - Scientific integrity (2021)	20.00
Seminars	
Thematic meeting: Radboudumc Technology Centers, Deep-Learning (2018)	1.50
Giving and receiving feedback (2019)	1.50
Conferentie Onderwijs en Onderzoek (2019)	1.00
Spiegeluur ORL on advances in science (2021)	78.00
Radboud Research Rounds (2021)	10.00

Conferences

PhD CaRe days (2018)	16.00
PhD retreat (2018)	16.00
7th Dutch Bio-Medical Engineering Conference (2019)	16.00
EORS 2019 (2019)	24.00
ICMS Event (2021)	1.00

Supervision

Supervision of Bachelor student internship - 6 months (2018)	80.00
Supervision of Bachelor student internship - 2 months (2019)	30.00
Supervision of Master student internship - 6 months (2019)	50.00
Supervision of Master student internship - 6 months (2021)	80.00
Supervision of Master student internship - 6 months (2023)	50.00
Total hours	**763.00**

CURRICULUM VITAE



Ali Ataei was born in Iran and developed an early interest in Biomedical Engineering and Biomechanics. He completed his bachelor's degree in Biomedical Engineering, with a specialization in Biomechanics, in 2009. He subsequently pursued an M.Sc. in Biomedical Engineering at Iran University of Science and Technology and graduated in 2012. His academic work during this period was primarily related to spinal biomechanics. Following his studies, Ali gained professional experience in the medical engineering industry, working with medical devices and technologies for pain management.

In January 2018, Ali moved to the Netherlands and started his PhD at the Orthopaedic Research Laboratory of Radboud University Medical Center in Nijmegen, under the supervision of Prof. Nico Verdonschot and Prof. Esther Tanck. His doctoral research focused on metastatic bone disease, combining biomedical imaging, computational modeling, and artificial intelligence. In particular, he investigated the use of deep neural networks for the automated detection and segmentation of metastatic lesions in the proximal femur from CT images. His research also explored characteristics of metastatic lesions and pathological bone in relation to the prediction of femoral strength and fracture risk.

Alongside the later stages of his doctoral research, Ali expanded his professional career into the medical device regulatory field. Since 2022, he has worked in medical device regulatory affairs, with his activities subsequently expanding to include quality assurance and clinical affairs. In these roles, he has combined his scientific and engineering background with regulatory and clinical aspects of medical device development and compliance.

ACKNOWLEDGEMENTS

Completing this thesis marks the end of an important chapter in my life. Looking back, I realize how many people contributed to this journey, both professionally and personally. I sincerely appreciate everyone who supported me, worked with me, shared their knowledge, or simply made these years more enjoyable.

First and foremost, I would like to express my sincere appreciation to my supervisors, **Prof. Nico Verdonschot** and **Prof. Esther Tanck**, for giving me the opportunity to pursue my PhD at the Orthopaedic Research Lab (ORL) and for their guidance throughout this journey. Nico, thank you for your scientific insight, your critical perspective, and for always encouraging me to look at the bigger picture. Esther, thank you for your thoughtful guidance, attention to details, and constructive feedback throughout the different stages of my research. I greatly appreciate the trust, freedom, and support you both gave me during these years. I would also like to thank my co-supervisors, **Dr. Nikolas Lessmann** and **Dr. Florieke Eggermont**, for sharing their expertise and contributing to different aspects of my research. Your discussions, suggestions, and scientific input helped me approach the challenges of this multidisciplinary project from different perspectives.

I would like to sincerely thank the members of my manuscript committee, **Prof. I. Sechopoulos, Dr. P.G. Westhoff, and Prof. C.I. Sánchez Gutiérrez**, for taking the time to read and evaluate my thesis. I greatly appreciate your willingness to be part of this important stage of my PhD and your valuable assessment of my work.

A large part of this journey took place at the Orthopaedic Research Lab. **Dennis**, although you were not officially one of my supervisors, you were always closely involved in the daily life of the lab. Thank you for always being approachable, helpful, and supportive throughout my years at the ORL. I would also like to thank **Richard, Martijn, Hans, Jasper, Thom, Esther, Erwin, Max, Corine, Thomas Anijs, Thomas Gersie, and Ineke** for making my time there memorable. I especially enjoyed the ORL activities and gatherings and, of course, our almost daily table-soccer games after lunch. Thank you all for the good times, laughter, and memories we shared over the years.

My PhD journey eventually overlapped with a new professional chapter in the medical device industry. I would like to thank my former colleagues at **Qserve** and my current colleagues at **TSC Life** for their support and flexibility while I was combining my professional responsibilities with completing my PhD. I appreciate the experiences we have shared and everything I have learned from working with you.

Beyond the ORL, I was fortunate to find wonderful friends in Nijmegen who made my life in the Netherlands much richer. **Matin**, thank you for your friendship and for all the fun and joyful moments we shared in Nijmegen. You made these years much more enjoyable, and I hope there are many more good times to come. I would also like to thank **Milad**, **Mona**, **Farid**, **Elnaz**, **Sadaf**, **Navid**, **Somayyeh**, **Fazael**, and **Alek** for their friendship and for the many good times we shared. Living far from home is very different when you are surrounded by good friends, and I feel fortunate that our paths crossed in the Netherlands. Thank you for the gatherings, celebrations, conversations, and laughter that gave me many wonderful memories beyond the PhD.

Most importantly, I want to thank my **parents** for their unconditional love, patience, and support throughout my life. Pursuing a PhD abroad meant being away from home for many years, but despite the distance, I always felt your presence and encouragement. Thank you for believing in me, for giving me the freedom and confidence to follow my own path, and for everything you have done and continue to do for me. For all of this and much more, thank you, **Maman** and **Baba**.

A very special thanks to **Nooshin** for her love, understanding, and patience, especially during the final stages of this journey. Her presence has made the difficult moments easier and the good moments even more enjoyable. I feel very lucky to have her in my life and to share this important moment with her.

I would also like to thank my sisters, **Sara** and **Saeideh**, for always being there for me and for remaining such an important part of my life despite the distance between us. Thank you for your love, encouragement, and all the moments we have shared as a family. I would also like to thank my brothers-in-law, **Mohammad** and **Shahin**, for their kindness and support and for being such an important part of our family.

Finally, I want to thank everyone who became part of this journey in one way or another. Your support, friendship, and encouragement made these years easier and more enjoyable, and I am happy to have shared this journey with so many wonderful people.

A


Radboudumc
university medical center
IN·DEI·NOMINE·FELICITER
Radboud University