User-centered co-development of Artificial Intelligence applications

Towards automated vertebral fracture assessment

Master Thesis | Technical Medicine

Marijn Mostert

Leiden University Medical Center

TU Delft

Erasmus MC University Medical Center Rotterdam

User-centered co-development of Artificial Intelligence applications

Towards automated vertebral fracture assessment

Jacob Marijn Mostert

Student number: 4373847

7th of July 2021

Thesis in partial fulfilment of the requirements for the joint degree of Master of Science in

Technical Medicine

Leiden University | Delft University of Technology | Erasmus University Rotterdam

Master thesis project (TM30004 | 35 ECTS)
Department of Radiology, Leiden University Medical Center
September 2020 – July 2021

Supervisors:

Dr. W. Grootjans, LUMC
Dr. L.M. Pereira Arias-Bouda, LUMC
Drs. D.D.D. Rietbergen, LUMC

Thesis committee members:

Prof. dr. ir. J. Harlaar, TU Delft (chair)
Dr. W. Grootjans, LUMC
Dr. L.M. Pereira Arias-Bouda, LUMC
Drs. D.D.D. Rietbergen, LUMC
Dr. L. Van Erven, LUMC

Summary

Vertebral fractures are the most common osteoporotic fractures, with a prevalence of 12-20% in Europe, making them a major health problem because of the associated morbidity, mortality and costs. Without adequate treatment, vertebral fractures are often followed by subsequent fractures, leading to further invalidation and deterioration of health. Therefore, early detection of vertebral fractures is important so preventative treatment can be initiated.

Vertebral fracture assessment (VFA) using Dual-Energy X-ray Absorptiometry (DXA) equipment is an imaging technique in which a lateral image of the spine is made. With vertebral morphometry, vertebral heights are measured and fractures are identified when vertebral height is lower than expected. However, vertebral morphometry is time and labor-intensive, and is subject to inter-operator variability. Automating VFA using Artificial Intelligence (AI) could help to overcome these limitations. We employed a co-development approach, aiming to create an AI-based tool to automatically perform vertebral morphometry on VFA images to identify vertebral fractures.

Firstly, we conducted a literature review to investigate the current use of AI in quantitative DXA imaging in a broad sense (Chapter 2). Besides VFA, other quantitative parameters describing bone macrogeometry and microgeometry can be extracted from DXA images. Incorporating these risk factors into multivariate prediction models could improve the identification of those at risk of fracture and help in clinical decision making. Although still in development, AI has been successfully applied in aid of fracture risk assessment, showing promising results.

In Chapter 3, the results of our reader study are described, evaluating VFA with manual vertebral morphometry as it is currently performed. This study served as a baseline measurement to quantify the effort needed to perform manual VFA and assess the potential value of automating VFA. The average annotation time per VFA image was 259 seconds. Although the intraclass correlation for vertebral height measurements between different readers was high, inter-observer agreement for fracture classification was only poor to moderate.

Together with our industry partner, we developed an AI-based software tool to perform vertebral morphometry. This tool is still in development, and we evaluated its current performance and potential impact in the study described in Chapter 4. Although its current standalone performance is suboptimal and shows room for improvement, this initial investigation showed that automated VFA has the potential to significantly reduce the required reader time.

Finally, in Chapter 5 we reflect on our user-centered co-development process and the next steps to bring our AI tool to market. We believe that collaboration between academic healthcare institutions and industry are essential for successful development of AI products. Validation of these products throughout the development process and actively involving intended users should be done. In the near future, vertebral fracture assessment can be supported by our AI-based application, potentially leading to lower annotation times and improving clinical workflow. However, further improvements have to be made and independent validation is required before market access.

Contents

1. General Introduction

1.1. Vertebral fractures

1.1.1. Epidemiology

Osteoporosis is defined as a skeletal disorder characterized by compromised bone strength predisposing a person to an increased risk of fracture [1]. Osteoporosis causes about 9 million fractures each year, and these osteoporotic fractures are a major health problem worldwide because of the associated morbidity, mortality and costs [2]. Vertebral fractures are the most common osteoporotic fractures, with a prevalence of 12-20% in Europe [3]. The risk of sustaining a vertebral fracture exponentially increases with age, due to the higher prevalence of osteoporosis in the elderly population, but other age-related factors such as risk of falls also play a role. The incidence of vertebral fractures is higher in women than in men [4], and prevalence is highest in postmenopausal women, with more than 50% of women aged 85 years or older having at least one vertebral fracture [5].

1.1.2. Clinical manifestation and impact

Vertebral fractures are often difficult to identify clinically, since symptoms are often absent or unspecific. They occur in the absence of trauma or during normal activities such as bending or turning, and as such the exact moment of onset is often unknown. Symptoms of vertebral fractures include back pain, limited spinal mobility, and loss of height [6]. Due to the reduced physical capability, vertebral fractures are associated with difficulty in performing activities of normal daily living and perceived poor general health, and as a result reduced independence and purposeful limitation of activity and social interactions. As such, vertebral fractures also have a major impact on patients' quality of life through pain, reduced physical capability, and emotional status [7].

1.1.3. Diagnosis

Despite their major impact on patients and society, vertebral fractures are difficult to identify clinically. Only about a third of vertebral fractures give clinical symptoms, and these symptoms are nonspecific and easily confused with other causes of back pain. Therefore, the diagnosis of a vertebral fracture mainly relies on radiographic assessment of the spine. Osteoporotic vertebral fractures are indicated by a collapse of the vertebral body (see Figure 1-1), and the associated alteration in the shape and reduction of height as a wedge, concave, or crush deformity [8].

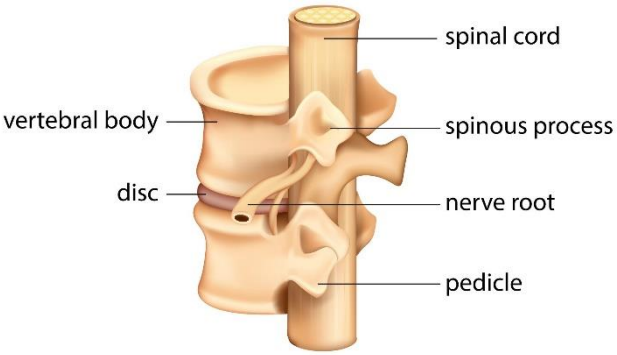


Figure 1-1. Basic anatomy of vertebrae. Adapted from [9]

Presence of a vertebral fracture, without major trauma or local disease, is a strong indicator for osteoporosis. In addition, a vertebral fracture is an independent predictor of subsequent osteoporotic fractures, not only in the spine but also the hip [10, 11]. Corrected for age and Bone Mineral Density (BMD), a vertebral fracture is associated with a four- to fivefold increase in risk of a subsequent vertebral fracture [11-13]. One in five osteoporotic women with a recent vertebral fracture will sustain a new vertebral fracture within the next year [14], and the risk of a new vertebral fracture increases with both the number and the severity of prevalent vertebral fractures [15-17]. Therefore, accurate diagnosis is of great importance to identify those at risk of subsequent fracture and to initiate treatment for the prevention of further deterioration of health.

1.1.4. Vertebral Fracture Assessment

Until recently, lateral spine radiographs were the most important diagnostic tool for the detection of vertebral fractures. However, densitometric vertebral fracture assessment (VFA) has gained popularity in the past decade. This imaging technique uses standard dual-energy X-ray absorptiometry (DXA) equipment that is also used for BMD measurement. This allows for measurement of two important fracture risk factors - BMD and vertebral fracture status - in one imaging session, reducing burden on patients and reducing costs [2, 18]. Sensitivity and specificity of VFA is adequate compared to radiographs [19, 20] and in addition, radiation exposure for VFA is much lower. A typical VFA results in a patient dose of 3 to 40 microSieverts (µSv), while the radiation exposure associated with conventional spinal radiography is around 600 µSv.

The International Society for Clinical Densitometry recommends performing VFA for all patients that fulfill the requirements described in Table 1-1 [21]. In addition, in a recent position paper, the International Osteoporosis Foundation (IOF) fracture working group advocated for the routine use of VFA in post-fracture care [22]. However, the detection of vertebral fractures on VFA images is not straightforward. Although every osteoporotic vertebral fracture is a vertebral deformity, not every vertebral deformity is a vertebral fracture. There is a lack of consensus about what height reduction constitutes a fracture, making VFA subject to significant inter-observer variability [23, 24]. Besides this, VFA requires trained interpreters to spend time identifying vertebral fractures, limiting the number of VFAs that can be performed during the day.

DXA equipment consists of a movable patient table and a C-arm with an x-ray source and detector, as depicted in Figure 1-2. To make a VFA image, patients are positioned in a lateral decubitus position or in supine position with the C-arm rotated 90 degrees. A lateral image of the thoracolumbar spine is made while the detector moves along the spine, minimizing parallax distortion effects that are present in conventional radiographs. With VFA, vertebrae from the fourth thoracic (T4) to the fourth lumbar (L4) vertebra are imaged. An example VFA image is shown in Figure 1-4.

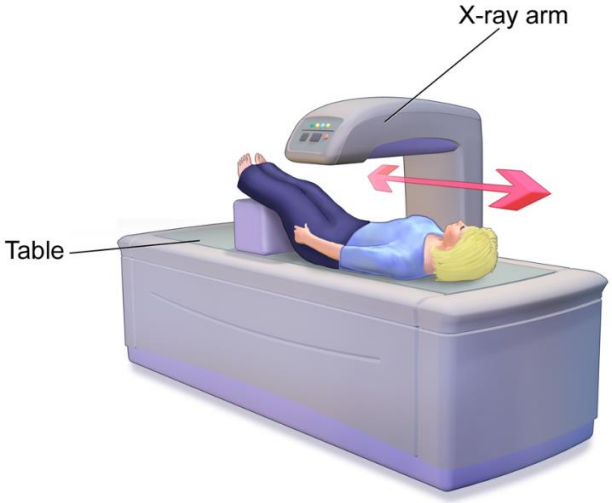


Figure 1-2. Schematic depiction of Dual-Energy X-ray Absorptiometry equipment used for Vertebral Fracture Assessment imaging. Adapted from [25].

Table 1-1. Indications for Vertebral Fracture Assessment according to the International Society for Clinical Densitometry [21].

T-score < -1.0 and one or more of the following:
Women > 70 years of age; Men > 80 years of age
Historical height loss > 4 cm
Self-reported prior vertebral fracture
Glucocorticoid therapy equivalent to ≥ 5 mg of prednisone per day for ≥ 3 months

Standardized methods for identifying vertebral fractures on VFA can be grouped in several categories. Algorithm-Based Qualitative (ABQ) methods consist of fully visual assessment using a series of standardized criteria, such as depression of the central vertebral endplate. In contrast, quantitative methods are based on measurement of vertebral shapes, also referred to as vertebral morphometry. There are also semi-quantitative methods that combine quantification of vertebral shape with visual inspection.

1.2. Vertebral Morphometry

Quantitative assessment methods were developed to create an objective and reproducible method for assessment of vertebral fractures, less dependent on the expertise of interpreters. Quantitative assessment generally consists of point placement along the vertebral body and subsequent measurement of vertebral heights. Based on these measurements, vertebrae are assigned a type and grade of fracture when vertebral heights are lower than expected.

Typically, six points are placed on each vertebra [8]. Four points are placed on the corners of each vertebral body from T4 to L4, and additional points in the middle of the upper and lower endplates. The uncinate process at the posterosuperior border of the thoracic vertebrae, Schmorl's nodes, and osteophytes should be excluded [26]. When the outer contours of the endplate are not superimposed (incorrect patient positioning or severe scoliosis), the middle points are placed in the center between the upper and the lower contour [8].

The Euclidean distance (d) from each point to the mid-vertebral line is calculated (see Figure 1-3) and used to determine the vertebral heights [27, 28]. The anterior height h_a, medial height h_m and posterior height h_p are given by:

$$h_a = d_E + d_F$$
$$h_m = d_c + d_D$$
$$h_p = d_A + d_B$$

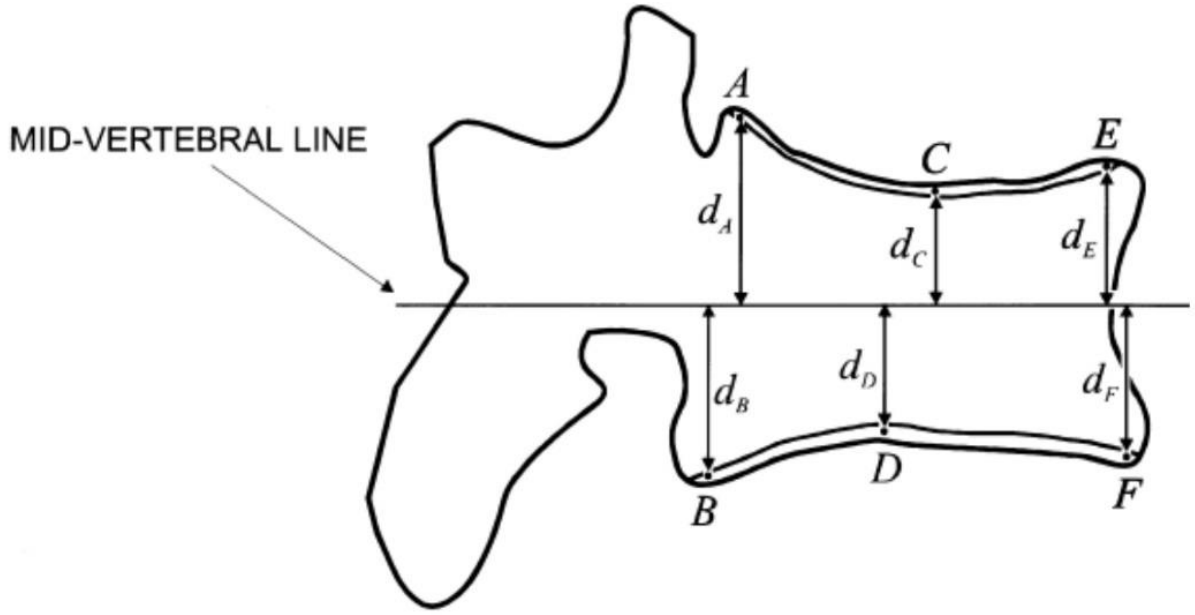


Figure 1-3. Six-point placement along a vertebral body. A & B posterior points, C & D middle points, and E & F anterior points. Distances from each point to the mid-vertebral line are used for the calculation of vertebral heights. Reprinted from [28].

As stated earlier, with quantitative assessment fractures are identified by vertebral heights that are lower than expected. There is however some normal variation in vertebral shape that does not constitute a fracture, complicating the definition of expected vertebral height. For example, mid-thoracic vertebrae can be slightly wedged, and lumbar vertebrae may have a biconcave shape [29]. Also, some normal changes in vertebral shape can be seen with age. Several thresholds for the extent of vertebral height reduction that constitutes a fracture have been proposed, based on comparison to other heights within the vertebra [30] or to population-based normative values [31-34]. Historically, a lot of debate has taken place about the best quantitative definition of vertebral fractures [35, 36]. Although there is no clear consensus on defining a threshold of height reduction that would allow unequivocal discrimination between vertebral fractures, deformities, and normal vertebrae, the currently most used definition uses a fixed percentage reduction based on the Genant classification scheme [23]. Height ratios are calculated according to Melton [37], but without normalizing to a reference population. The wedge ratio is calculated by dividing anterior height by posterior height (h_a/h_p), biconcavity is calculated by dividing mid-height by posterior height (h_m/h_p), and the crush ratio is determined by dividing posterior height by posterior heights of adjacent vertebrae. Fracture gradation is done using the thresholds as defined by Genant [38]: normal for height loss less than 20%, mild fracture (grade 1) for height loss between 20% and 25%, moderate fracture (grade 2) for ≥25% and <40%, and severe fracture (grade 3) for height reduction more than 40%. The different fracture grades and types are schematically shown in Figure 1-5.

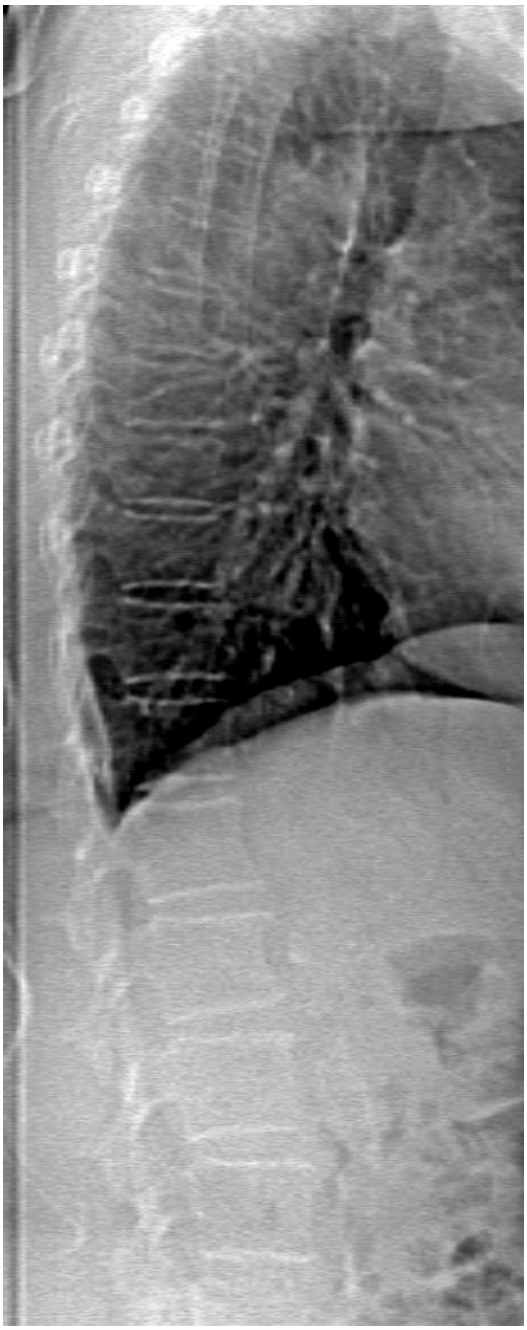


Figure 1-4. Example Vertebral Fracture Assessment Image

1.2.1. Treatment

Clinical management of vertebral fractures is generally focused on treatment of symptoms. Possible treatment strategies include pharmacological pain treatment or local steroid injections, physical therapy, and bracing [6]. In severe cases with debilitating, treatment refractory pain, vertebroplasty or kyphoplasty can be considered, in which the fractured vertebral body is surgically injected with a polymerous bone cement [39]. However, most vertebral fractures heal without surgical intervention and pain decreases over time. In addition, treatment of osteoporotic fractures should also focus on preventing incidence of new fractures and progression of existing fractures [6]. Medicinal treatment can be divided into two categories. Anti-resorptive agents such as estrogen and bisphosphonates reduce bone resorption. Anabolic agents on the other hand stimulate bone formation. Therapeutic decisions regarding the treatment of choice are based on individual patients' risk of developing additional fractures.

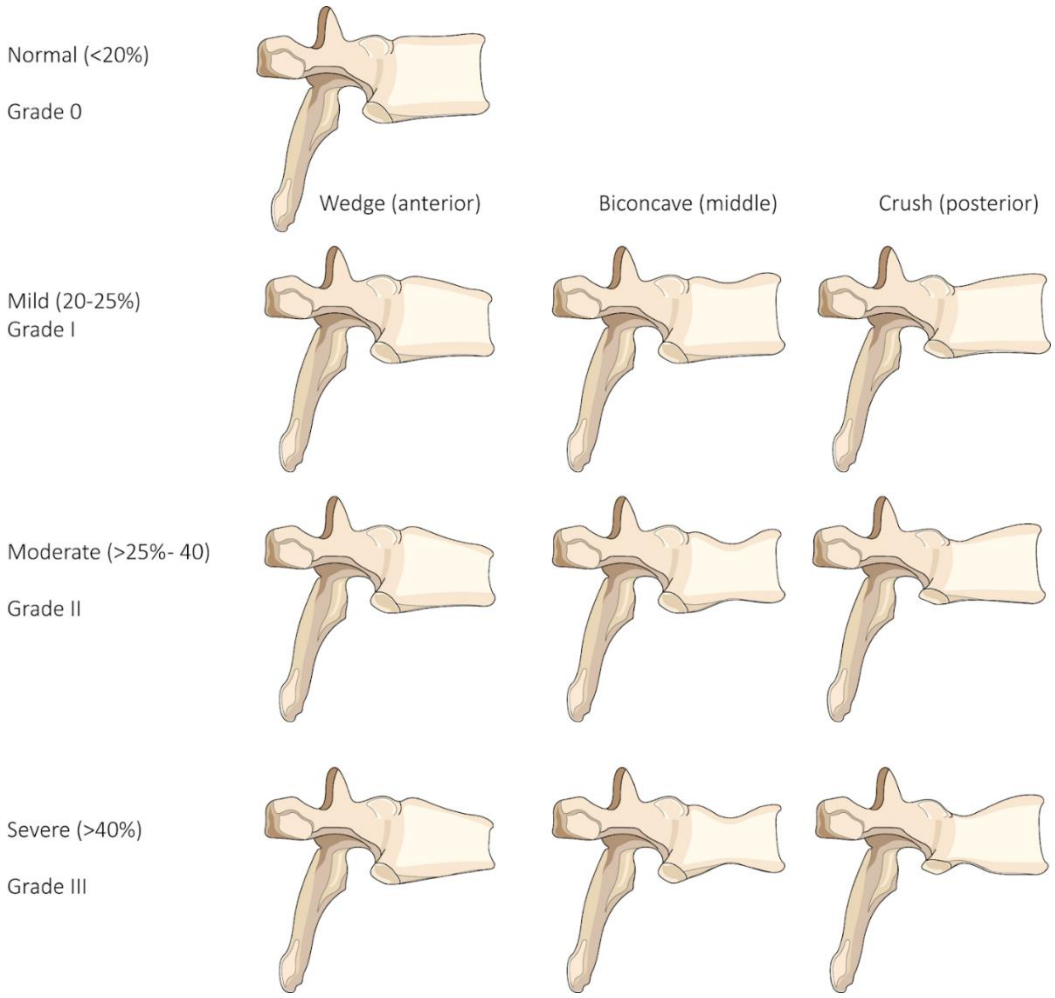


Figure 1-5. Genant classification of normal and fractured vertebrae with different fracture types and severities.

1.3. Artificial Intelligence

The term Artificial Intelligence (AI) was coined in 1956 by computer scientist John McCarthy to describe the field of research aiming to create machines that are able to perform tasks that generally require some form of intelligence [40]. Early AI methods led to applications with varying degrees of utility and mostly subhuman performance. However, with recent advances in theoretical understanding, increased computer power, and availability of large amounts of data, AI use is currently on an upward trend [41]. Similarly in medicine, AI research is becoming more and more widespread. The field of radiology, heavily relying on imaging data, is at the forefront of this development, and over 100 AI products for radiology are already commercially available [42].

The term AI encompasses a very broad set of methods, from simple rule-based expert systems to advanced black-box algorithms. However, there are two subsets of AI methods that involve some form of learning from examples, which make up the majority of AI applications in medicine [44]. Conventional Machine Learning uses deterministic features that can be defined as mathematical equations, which are engineered by humans based on expert knowledge and require specification of a region or volume of interest. Examples of features are shape, size and texture. These features are then used to train machine learning models to classify patients, and these models can then be used to assist in clinical decision making. Artificial Neural Networks (ANN) are a subset of machine learning models that resemble information processing in the human brain. In its most basic form, ANNs consist of an interconnected group of individual units called neurons. These neurons are generally organized in

multiple layers, with an input layer, output layer, and one or more hidden layers in between. ANNs with more than one hidden layer are often referred to as Deep Learning models.

The main advantage of deep learning is that it does not require manually selected features to be trained on, thus factoring out the need for expert knowledge and the possible subjectivity that comes with manual feature selection. Instead, deep learning algorithms can automatically learn feature representations from data, allowing for more abstract feature definitions. In direct comparisons, deep learning methods often outperform other machine learning methods [45]. When working with image data, as is common in radiology, a model architecture called a Convolutional Neural Network (CNN) is most often employed [46]. CNNs consist of convolutional layers, pooling layers, and a fully connected layer, each with distinct functions (see Figure 1-6). Convolutional layers work as spatial filters, identifying certain image characteristics, and pooling layers reduce dimensionality by selecting which features are maintained. By repeating convolutional and pooling layers, increasingly higher-level features can be extracted. Earlier layers might learn abstract shapes such as lines and contrasts, while deep layers might learn entire objects or organs.

Finally, fully connected layers can use the extracted information to classify images, for example for the differentiation between benign and malignant tumors. Besides classification, other deep learning architectures may learn to segment regions of interest, perform measurements, or solve regression problems such as determining bone age [32, 47]. In essence, deep learning algorithms mimic trained radiologists in identifying image parameters and weighing the importance of these parameters to make clinical decisions.

1.4. Research Goals

This thesis explores the co-creation of an AI-based application for automated detection of vertebral fractures on VFA, as part of a larger project aimed at improving radiology through value-based innovation. The main research question is whether automated VFA is viable and can add value to clinical practice. Firstly, the use of quantitative DXA measurements is reviewed in a broad sense, including VFA and other quantitative parameters regarding patients' risk of developing osteoporotic fractures. Then, the currently used method involving manual vertebral morphometry for fracture detection on VFA is evaluated, focusing on inter-observer agreement and reader effort. Finally, the performance and potential added value of an AI algorithm to automatically perform vertebral morphometry on VFA images is evaluated, as well as its co-development process.

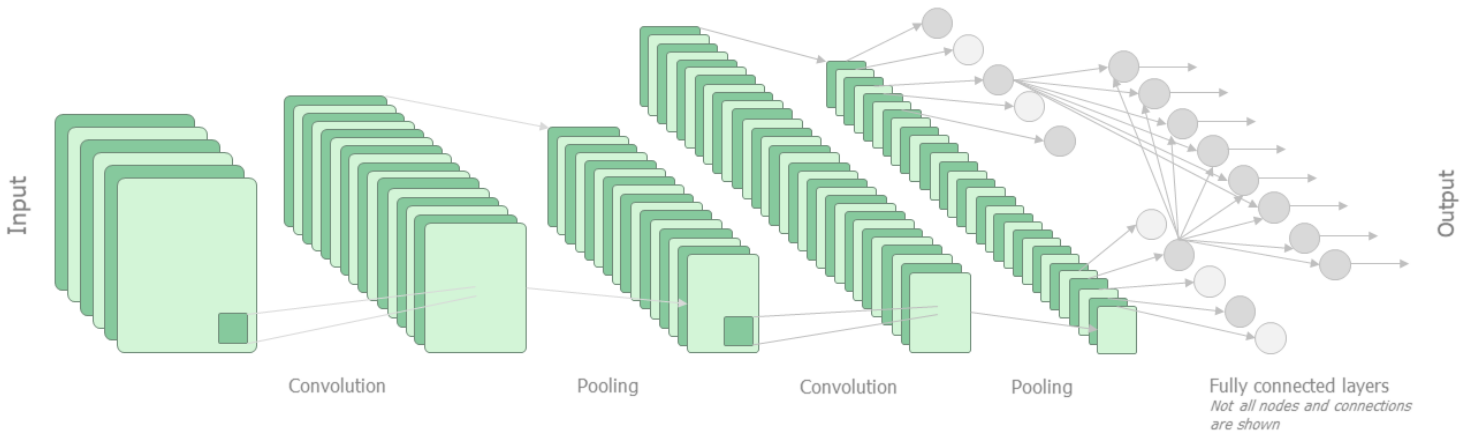


Figure 1-6. Schematic representation of a Convolutional Neural Network.

1.5. References

1. Lorentzon M, Cummings SR (2015). Osteoporosis: the evolution of a diagnosis. J Intern Med. 277(6):650-661. https://doi.org/10.1111/joim.12369
2. Schousboe JT, Vokes T, Binkley N, Genant HK, Sambrook P, Pocock N (2010). Vertebral Fracture Initiative Part 3: Densitometric vertebral fracture assessment (VFA). International Osteoporosis Foundation 12:1-24.
3. O'neill TW, Felsenberg D, Varlow J, Cooper C, Kanis JA, Silman AJ, European Vertebral Osteoporosis Study Group (1996). The prevalence of vertebral deformity in European men and women: the European Vertebral Osteoporosis Study. J Bone Miner Res. 11(7):1010-1018.
4. Felsenberg D, Silman AJ, Lunt M et al (2002). Incidence of vertebral fracture in europe: results from the European Prospective Osteoporosis Study (EPOS). J Bone Miner Res. 17(4):716-724.
5. Diacinti D, Guglielmi G (2010). Vertebral morphometry. Radiol Clin North Am. 48(3):561-575. https://doi.org/10.1016/j.rcl.2010.02.018
6. Szule P, Bouxsein ML (2011). Vertebral Fracture Initiative Part I, Overview of osteoporosis: Epidemiology and clinical management. International Osteoporosis Foundation.
7. Ross PD, Ettinger B, Davis JW, Melton L, Wasnich RD (1991). Evaluation of adverse health outcomes associated with vertebral fractures. Osteoporos Int. 1(3):134-140.
8. Guglielmi G, Diacinti D, Van Kuijk C, Aparisi F, Krestan C, Adams JE, Link, TM (2008). Vertebral morphometry: current methods and recent advances. Eur Radiol. 18(7):1484-1496.
9. Human Vertebrae Anatomy (online). Available on https://www.freepik.com
10. Melton LJ, Atkinson EJ, Cooper C, O'Fallon WM, Riggs BL (1999). Vertebral fractures predict subsequent fractures. Osteoporos Int. 10:214-221.
11. Black DM, Arden NK, Palermo L, Pearson J, Cummings SR (1999). Prevalent vertebral deformities predict hip fractures and new vertebral deformities but not wrist fractures. Study of Osteoporotic Fractures Research Group. J Bone Miner Res 14(5):821-828.
12. Cauley JA, Hochberg MC, Lui LY, Palermo L, Ensrud KE, Hillier TA, Nevitt MC, Cummings SR (2007). Long-term risk of incident vertebral fractures. JAMA 298(23):2761-2767.
13. Lindsay R, Silverman SL, Cooper C, Hanley DA, Barton I, Broy SB, Licata A, Benhamou L, Geusens P, Flowers K, Stracke H, Seeman E (2001). Risk of new vertebral fracture in the year following a fracture. JAMA. 285(3):320-323.
14. Johnell O, Oden A, Caulin F, Kanis JA (2001). Acute and long-term increase in fracture risk after hospitalization for vertebral fracture. Osteoporos Int. 12:207-214.
15. Nevitt MC, Ross PD, Palermo L, Musliner T, Genant HK, Thompson DE (1999). Association of prevalent vertebral fractures, bone density, and alendronate treatment with incident vertebral fractures: effect of number and spinal location of fractures. Bone 25:613-619.
16. Bouxsein ML, Chen P, Glass EV, Kallmes DF, Delmas PD, Mitlak BH (2009). Teriparatide and raloxifene reduce the risk of new adjacent vertebral fractures in postmenopausal women with osteoporosis. Results from two randomized controlled trials. J Bone Joint Surg Am. 91:1329-1338.
17. Siris ES, Genant HK, Laster AJ, Chen P, Misurski DA, Krege JH (2007). Enhanced prediction of fracture risk combining vertebral fracture status and BMD. Osteoporos Int. 18:761-770.
18. Shetty S, John B, Mohan S, Paul TV (2020). Vertebral fracture assessment by dual-energy X-ray absorptiometry along with bone mineral density in the evaluation of postmenopausal osteoporosis. Arch Osteoporos. 15(1):1-6.
19. Malgo F, Hamdy NAT, Ticheler CHJM, Smit F, Kroon HM, Rabelink TJ, Dekkers OM, Appelman-Dijkstra NM (2017). Value and potential limitations of vertebral fracture assessment (VFA) compared to conventional spine radiography: experience from a fracture liaison service (FLS) and a meta-analysis. Osteoporos Int. 28(10):2955-2965.

20. Lee JH, Lee YK, Oh SH, Ahn J, Lee YE, Pyo JH, Choi YY, Kim D, Bae SC, Sung YK, Kim DY (2016). A systematic review of diagnostic accuracy of vertebral fracture assessment (VFA) in postmenopausal women and elderly men. Osteoporos Int. 27(5):1691-1699.
21. Shuhart CR, Yeap SS, Anderson PA, Jankowski LG, Lewiecki EM, Morse LR, Rosen HN, Weber DR, Zemel BS, Shepherd JA (2019). Executive summary of the 2019 ISCD position development conference on monitoring treatment, DXA cross-calibration and least significant change, spinal cord injury, peri-prosthetic and orthopedic bone health, transgender medicine, and pediatrics. J Clin Densitom. 22(4):453-471.
22. Lems WF, Paccou J, Zhang J, Fuggle NR, Chandran M, Harvey NC, Cooper C, Javaid K, Ferrari S, Akesson KE (2021). Vertebral fracture: epidemiology, impact and use of DXA vertebral fracture assessment in fracture liaison services. Osteoporos Int. 32(3):399-411. https://doi.org/10.1007/s00198-020-05804-3
23. Oei L, Koromani F, Breda SJ et al (2018). Osteoporotic vertebral fracture prevalence varies widely between qualitative and quantitative radiological assessment methods: the Rotterdam Study. J Bone Miner Res. 33(4):560-568.
24. Lentle B, Koromani F, Brown JP et al (2019). The radiology of osteoporotic vertebral fractures revisited. J Bone Miner Res. 34(3):409-418. https://doi.org/10.1002/jbmr.3669
25. Blausen.com staff (2014). Medical gallery of Blausen Medical 2014. WikiJournal of Medicine 1 (2). https://doi.org/10.15347/wjm/2014.010
26. Hurxthal LM (1968) Measurement of anterior vertebral compressions and biconcave vertebrae. AJR Am J Roentgenol 103(3):635-644.
27. Blake GM, Rea JA, Fogelman I (1997) Vertebral morphometry studies using dual-energy x-ray absorptiometry. Semin Nucl Med 27(3):276-290.
28. Harvey SB, Hutchison KM, Rennie EC, Hukins DW, Reid DM (1998). Comparison of the precision of two vertebral morphometry programs for the lunar EXPERT-XL imaging densitometer. Br J Radiol. 71(844):388-398.
29. Grigoryan M, Guermazi A, Roemer FW, Delmas PD, Genant HK (2005). Recognizing and reporting osteoporotic vertebral fractures. The Aging Spine 22-30.
30. Melton LJ, Kan SH, Frye MA, Wahner HW, O'fallon WM, Riggs BL (1989). Epidemiology of vertebral fractures in women. Am J Epidemiol. 129(5):1000-1011.
31. McCloskey EV, Spector TD, Eyres KS, Fern ED, O'rourke N, Vasikaran S, Kanis JA (1993). The assessment of vertebral deformity: a method for use in population studies and clinical trials. Osteoporos Int. 3(3):138-147.
32. Eastell, R, Cedel SL, Wahner HW, Riggs BL, Melton LJ (1991). Classification of vertebral fractures. J Bone Miner Res. 6(3):207-215.
33. Black DM, Cummings SR, Stone K, Hudes E, Palermo L, Steiger P (1991). A new approach to defining normal vertebral dimensions. J Bone Miner Res. 6(8):883-892.
34. Minne HW, Leidig G, Wüster CHR, Siromachkostov L, Baldauf G, Bickel R, Sauer P, Lojen M, Ziegler R (1988). A newly developed spine deformity index (SDI) to quantitate vertebral crush fractures in patients with osteoporosis. Bone Miner. 3(4):335-349.
35. Grados F, Roux C, De Vernejoul MC, Utard G, Sebert JL, Fardellone P (2001). Comparison of four morphometric definitions and a semiquantitative consensus reading for assessing prevalent vertebral fractures. Osteoporos Int. 12(9):716-722.
36. Black DM, Palermo L, Nevitt MC, Genant HK, Christensen L, Cummings SR, Study of Osteoporotic Fractures Research Group (1999). Defining incident vertebral deformity: a prospective comparison of several approaches. J Bone Miner Res. 14(1):90-101.
37. Melton LJ, Kan SH, Frye MA, Wahner HW, O'fallon WM, Riggs BL (1989). Epidemiology of vertebral fractures in women. Am J Epidemiol. 129(5):1000-1011.

38. Genant HK, Wu CY, Van Kuijk C, Nevitt MC (1993). Vertebral fracture assessment using a semiquantitative technique. J Bone Miner Res. 8(9):1137-1148.
39. Watts NB, Harris ST, Genant HK (2001). Treatment of painful osteoporotic vertebral fractures with percutaneous vertebroplasty or kyphoplasty. Osteoporos Int. 12:429-437.
40. McCarthy J (1988). Review of The Question of Artificial Intelligence. Ann Hist Comput. 10 (3):224–229.
41. Grace K, Salvatier J, Dafoe A, Zhang B, Evans O (2018). When will AI exceed human performance? Evidence from AI experts. J Artif Intell Res. 62:729-754.
42. Van Leeuwen KG, Schalekamp S, Rutten MJ, Van Ginneken B, De Rooij M. (2021). Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. Eur Radiol. Apr 15:1-8.
43. Fujita H (2020). AI-based computer-aided diagnosis (AI-CAD): the latest review to read first. Radiol Phys Technol. 13(1):6-19.
44. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJ (2018) Artificial intelligence in radiology. Nat Rev Cancer 18(8):500-510.
45. Currie G, Hawk KE, Rohren E, Vial A, Klein R (2019). Machine learning and deep learning in medical imaging: intelligent imaging. J Med Imaging Radiat Sci. 50(4):477-487. https://doi.org/10.1016/j.jmir.2019.09.005
46. Gao J, Jiang Q, Zhou B, Chen D (2019). Convolutional neural networks for computer-aided detection or diagnosis in medical image analysis: an overview. Math Biosci Eng. 16(6):6536-6561.
47. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJ (2018). Artificial intelligence in radiology. Nat Rev Cancer. 18(8):500-510.

2. Fracture risk assessment using Artificial Intelligence and Dual Energy X-ray Absorptiometry: a Literature Review

In collaboration with D.D.D. Rietbergen, L.M. Pereira Arias-Bouda, and W. Grootjans

2.1. Abstract

Osteoporotic fractures are a major health problem worldwide because of the associated morbidity, mortality and costs. Early identification of individuals at risk of osteoporotic fracture is vital in the prevention strategy, so treatment can be initiated in a timely manner. Dual Energy X-ray Absorptiometry (DXA) is used for measurement of Bone Mineral Density, but its quantitative nature makes this modality suitable for other quantitative measurements such as Trabecular Bone Score, Hip geometry, and vertebral fracture assessment. Advanced computational methods based on Artificial Intelligence have been developed to support fracture risk assessment based on DXA imaging. In this article we review the different indicators of fracture risk that can be determined using DXA imaging, the use of Artificial Intelligence to determine these risk factors, and their value in fracture risk assessment.

2.2. Introduction

Osteoporosis is a skeletal disorder characterized by compromised bone strength, and consequently increased risk of fracture [1]. As osteoporosis is often asymptomatic, it may only come to light after a patient presents with a fracture without major trauma [2]. The prevalence of osteoporosis is higher in women and increases with age, from 19% among women aged 65 to 74 years to >50% in women aged ≥85 years [3]. Worldwide, osteoporosis causes more than 8.9 million fractures annually [4], and 1 in 3 women over age 50 will experience osteoporotic fractures, as will 1 in 5 men aged over 50 [5 - 7]. With the increasingly aging population, the prevalence of osteoporosis is only expected to increase.

Osteoporotic fractures are a major health problem worldwide because of the associated morbidity, mortality and costs. Hip and vertebral fractures are the most important types of osteoporotic fracture, since hip fractures are associated with a substantial increase in risk of institutionalization and death [8], and vertebral fractures are associated with chronic back pain, spinal deformity, functional limitations, and increased risk of hospitalization and mortality [9]. The mortality rate five years after a hip or spine fracture is about 20% greater than expected [10]. In addition, many patients who have an osteoporotic fracture lose the ability to perform normal activities of daily living, which may negatively affect self-esteem and emotional status and lead to depression [11, 12]. Adequate treatment can help prevent osteoporotic fractures and all its comorbidities and healthcare costs. Therefore, early identification of individuals at risk of osteoporotic fracture is vital in the prevention strategy, so treatment can be initiated in a timely manner. Furthermore, patients at high risk of fracture may also require more intensive treatment than low-risk patients, making fracture risk an important parameter in clinical decision making.

Assessment of fracture risk is based on clinical risk factors and measurement of Bone Mineral Density (BMD) using Dual Energy X-ray Absorptiometry (DXA). DXA is a fast, noninvasive, and reliable imaging method for measurement of BMD, and BMD measured by DXA became the reference standard for diagnosis of osteoporosis [13]. The quantitative nature of DXA makes this modality also suitable for

other quantitative measurements beyond purely visual assessment. Beside BMD, DXA can be used to determine indicators of bone microarchitecture, bone geometry, and presence of fracture, which can be used to refine fracture risk assessment [14]. Having recently gained traction, these features can be extracted from DXA images used for BMD measurement or from additional DXA images.

Advanced computational methods based on Artificial Intelligence (AI) have been developed to support the fracture risk assessment based on DXA imaging. The term AI is used to describe computer systems that are able to perform tasks that normally require human intelligence. Machine learning and deep learning are subsets of AI, in which a computer is trained to perform a task by supplying it with training data. AI can be used for a large number of applications. In this article we review the different indicators of fracture risk that can be determined using DXA imaging, the use of AI to determine these risk factors, and their value in fracture risk assessment.

2.3. Bone Mineral Density Measurement

2.3.1. Indications

Measurement of BMD is the most important technique for making the diagnosis for osteoporosis. In fact, osteoporosis is clinically defined as a reduction in bone mass as measured by DXA [15]. DXA is considered the golden standard for BMD measurement, and is therefore used as the reference standard to quantify osteoporosis. The International Society for Clinical Densitometry set out the indications for BMD measurement in their official guidelines [13] (see Table 2-1). As mentioned, the main application of BMD is the diagnosis of osteoporosis, but it may additionally be used to predict the risk of fractures, assist in therapeutic decision making, and help to monitor treatment response.

Table 2-1. Indications for Bone Mineral Density measurement with DXA according to the International Society for Clinical Densitometry [13].

Women	Aged 65 years and older Younger than 65, peri- or post-menopausal, with risk-factors for low bone mass
Men	Aged 70 and older Younger than 70 with risk factors for low bone mass
Both	Adults with a fragility fracture Adults with a disease or condition associated with low bone mass or bone loss Adults taking medications associated with low bone mass or bone loss Anyone being considered for pharmacologic therapy Anyone being treated, to monitor treatment effect Anyone not receiving therapy in whom evidence of bone loss would lead to treatment

2.3.2. Measurement Locations

Measurement of BMD is possible in a number of anatomical locations. Axial DXA of the lumbar spine and proximal femur are the preferred techniques, because of their good reliability and precision [16, 17]. The spine contains a high proportion of trabecular bone, so spinal BMD measurement allows for early detection of high bone turnover. Spinal BMD is measured using a postero-anterior scan of the lumbar spine from L1 to L4. The mean BMD of these vertebrae is determined, after exclusion of vertebrae with fractures or focal lesions (e.g. metastatic, osteoblastic bone lesions) [16]. Measurement of BMD of the hip can be done on the total proximal femur or the femoral neck alone, usually whichever has the lower BMD [17]. The accuracy of axial bone densitometry is high, with a margin of error of 1–2% [18].

2.3.3. Measured parameters

The amount of X-ray energy that is attenuated by bone mineral in a specific region of interest determines the measured bone mineral content (BMC). BMD is determined by dividing BMC by the area of the region of interest. Since BMC is measured in a two-dimensional projection of the bone, this method measures areal BMD (aBMD) in g/cm^2. Three-dimensional imaging techniques are able to determine volumetric BMD, but this is not possible with conventional DXA equipment.

For diagnosis of osteoporosis, BMD measurements are usually expressed as T-scores or Z-scores, comparing measured values to a population-derived normative value. The T-score is used to diagnose osteoporosis in postmenopausal women and in men aged 50 and over. The T-score is the number of standard deviations the measured BMD deviates from the mean of a young adult reference population of the same sex. T-scores of -2.5 or less are considered osteoporosis, scores above -1 are considered normal, and score between -1 and -2.5 are in the osteopenic range [13]. For pre-menopausal women and men younger than 50 years, Z-scores are used. Z-scores are determined by comparing a patient's BMD to the mean for a reference population of the same age, race, and sex.

2.3.4. Fracture prediction

BMD measured with DXA can be used to estimate fracture risk. Because of the systematic nature of osteoporosis, low BMD in one site is associated with increased fracture risk across other skeletal sites. For every reduction in BMD of 1 standard deviation (SD), overall fracture risk increases by 1.5 - 2 times [19, 20]. Hip BMD may be somewhat more strongly related to most fracture types than spine BMD, especially for hip fracture, with an increase in risk ratio of approximately 3 for each SD decrease in BMD.

However, there is a wide overlap in the bone densities of patients who develop a fracture and those who do not. BMD accounts for only 60–70% of the variation in bone strength [21], and is only one of a number of important risk factors for fracture [14]. In fact, most fragility fractures occur in individuals with BMD values in the osteopenic or normal range [2, 19, 22]. Consequently, BMD measurement alone has a low sensitivity when used for osteoporosis screening [14]. Although BMD alone is not sufficient to identify individuals who will have a fracture, BMD measurements can identify people who have an increased risk of developing a fracture and can be a useful risk assessment parameter when used in combination with other clinical risk factors, such as age, Body Mass Index, and corticosteroid use.

2.3.5. Automated segmentation

For reliable measurement of BMD, accurate selection of the Region of Interest (ROI) in which BMD is determined is of great importance. Currently, ROI selection is usually done manually. With manual ROI selection, substantial variability can be induced, and research has shown that BMD measurement is highly sensitive to ROI selection and minor adjustments significantly affect measured BMD [23 - 27]. To overcome this problem, Hussein & Han (2019) proposed a computer assisted method for osteoporosis diagnosis using a machine learning algorithm [28]. First, a pixel label random forest model was used to distinguish bone from soft tissue. After this segmentation step, body-site specific algorithms were used to select ROIs (lumbar spine in both anterior-posterior and lateral images, proximal femur, and forearm interosseous membrane). Using this method to measure BMD on three consecutive spine phantom scans, the computer assisted method showed an average standard deviation of 0.0029, compared to 0.1199 for manual measurements by three different individuals. Similar results were found for femur BMD in a human subject. Although this fully automatic BMD measurement method showed promising results, independent validation in a larger population is needed.

Lateral spine DXA images, which are usually made for VFA purposes, can also show vascular calcifications in the thoracolumbar aorta [29, 30]. Calcification scores derived from these images are

associated with the risk of cardiovascular events [31, 32]. However, for postero-anterior BMD measurement, vascular calcifications are a potential source of error, as the calcifications are projected onto vertebrae and artificially increase measured BMD values. To acquire the most accurate BMD measurements, vertebrae affected by these vascular calcifications or other artefacts such as osteophytes should be excluded from the analysis. This can be done manually, or by a specialized algorithm. Tsang & Leslie compared three computer algorithms to manual exclusion of vertebrae and observed fair to moderate agreement between physicians and computer methods [33]. All methods of vertebral exclusion led to a small improvement in fracture prediction, and the automated algorithms performed at least as well as physicians when fractures were used as the endpoint. These results indicate that automated methods of vertebral exclusion can provide an objective counterpart to the subjective manual method.

2.4. Trabecular bone score

BMD is a measure of bone density, and as such an indicator of bone strength, but it does not say anything about the quality of the bone. The quality of bone microarchitecture is an important determinant of bone strength, independent of BMD [34, 35]. The Trabecular Bone Score (TBS) was therefore proposed as a texture parameter that reflects pixel gray-level variations in DXA images [36]. TBS is derived from the slope of the experimental variogram within a region of interest. In other words, TBS is a measure of how much pixel values vary depending on the distance between them (see Figure 2-1). A tight network of trabeculae produces a two-dimensional projection image with many gray-level variations of small amplitude and therefore a high TBS value, associated with good mechanical strength. A low TBS value, in contrast, indicates poor-quality microarchitecture with few gray-level variations of considerable amplitude.

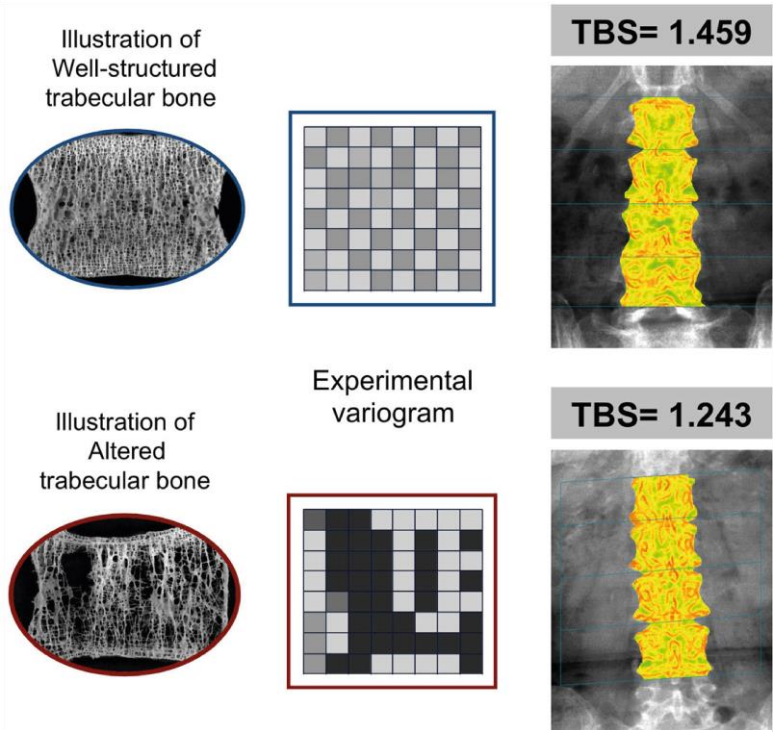


Figure 2-1. TBS principles. Densely interconnected trabecular networks translate into a high TBS value, whereas altered trabecular bone with high trabecular separation results in a lower TBS. Adapted from [37]. TBS: Trabecular Bone Score.

TBS is generally derived from the lumbar spine DXA image, and can be performed simultaneously with BMD measurement. Vertebrae excluded from BMD analysis are also excluded from TBS calculation. For postmenopausal women, a normal range for TBS values has been proposed: TBS above 1.350 is considered to be normal; TBS between 1.200 and 1.350 is considered partially degraded microarchitecture; and TBS below 1.200 defines degraded microarchitecture [37]. A normal range for TBS in men has not yet been proposed.

Evaluation of TBS in ex-vivo studies has shown significant correlations between TBS and various micro-CT derived parameters [38 - 40], but other ex-vivo studies have not found any significant correlations between TBS and microarchitecture [41, 42]. In-vivo studies have found weak to moderate correlations

[43 - 45]. The relation between TBS and trabecular architecture is not yet fully established, and remains subject of investigation [46].

It is worth noting that bone microarchitecture cannot be directly measured with DXA, as its resolution is not good enough to resolve bone trabeculae [47]. It is possible that TBS captures macroscopic features that correlate with microarchitecture, but this also remains subject of debate. Nevertheless, in the Manitoba cohort consisting of men and women aged 40 years or older, TBS was significantly lower in patients with osteoporosis, incident fractures, COPD, diabetes, high alcohol use, prior fracture, glucocorticoid use, and in women. Conversely, recent osteoporosis therapy was associated with increased TBS [48].

TBS also seems to be able to identify patients at risk of fracture independently from BMD and other predictors. In a recent meta-analysis comprised of 17,809 patients with a mean follow-up of 6.7 years, TBS showed an overall gradient of risk (GR; hazard ratio per 1 SD change in risk variable in direction of increased risk) of 1.44 (95% confidence interval (CI) 1.35–1.53) for major osteoporotic fracture when adjusted for age and time since baseline [49]. An important characteristic of a risk factor is that it is independent of other risk factors that are possibly easier to acquire. In the same meta-analysis, TBS was adjusted for BMD and commonly used clinical risk factors, and still remained a significant predictor of major osteoporotic fracture (GR=1.32, 95% CI 1.24–1.41). This suggests that TBS is indeed a promising independent risk factor.

2.5. Hip geometry

From a mechanical perspective, it is clear that the risk of fracture of an object is not solely determined by its material properties, but also by its load determined by geometry, both on macro- and microscopic level. In fact, the reason why areal BMD is so strongly correlated with fracture risk is partly because aBMD values include a measure of both bone mineralization and bone size. A small bone with a high volumetric BMD can be significantly weaker than a large bone with a smaller volumetric BMD, even though they have the same aBMD [50]. As such, aBMD alone is not a complete characterization of bone strength.

In the early 1990's, the first attempts were made to capture geometric properties of bone from DXA scans and include these in the fracture assessment of the hip [51]. This method, termed Hip Strength Analysis (HSA), was based on characterization of the projection profile extracted from planar DXA along the proximal femur at the narrow neck, intertrochanter, and shaft regions. Eight parameters per site are calculated which together provide a comprehensive structural bone assessment. These are bone resistance to bending (cross-sectional moment of inertia and section modulus), compression (cross-sectional area), and buckling (buckling ratio), combined with bone size (bone width, endosteal diameter, cortical thickness, aBMD) [52]. However, these eight parameters are not independent [53] and clinical studies evaluating them, either individually or combined, were unable to show improved predictive value compared to aBMD alone [54 - 57].

Khoo et al. later proposed a simplified base measurement framework that captures the same information in four parameters that was conventionally reported using the 8 measures [53] . These include aBMD and bone width (W), equivalent to in the HSA method described earlier, with the addition of a measure of the displacement (δ) between the geometric centre of the mineral-mass projection profile and its centre-of-mass, and a measure of the standard deviation of the mineral-mass projection profile (σ), as depicted in Figure 2-2. Predictive value of these measurements, combined with age and total aBMD of the proximal femur were evaluated using stepwise logistic regression models [58]. At all three sites, the σ representing the spread of bone mass within the section, was associated with the

development of hip fracture. Bone width was associated with fracture at the narrow neck and the shaft, and δ was increased in participants with hip fracture only at the narrow neck. Logistic regression with total hip aBMD, age, and σ at the intertrochanter region reached the best predictive performance, with an area under the receiver operator characteristic curve of 0.73, significantly higher than age and aBMD alone (0.69, p=0.009).

This seems to indicate that σ, especially at the intertrochanter region, can be seen as an additional indicator of fracture risk. While aBMD captures information regarding the amount of mineral within a section of the bone, σ adds information regarding the spatial distribution of the mineral within the measured section. As such, σ is influenced by multiple factors such as overall bone size, cortical thickness and the ratio between cortical and trabecular bone.

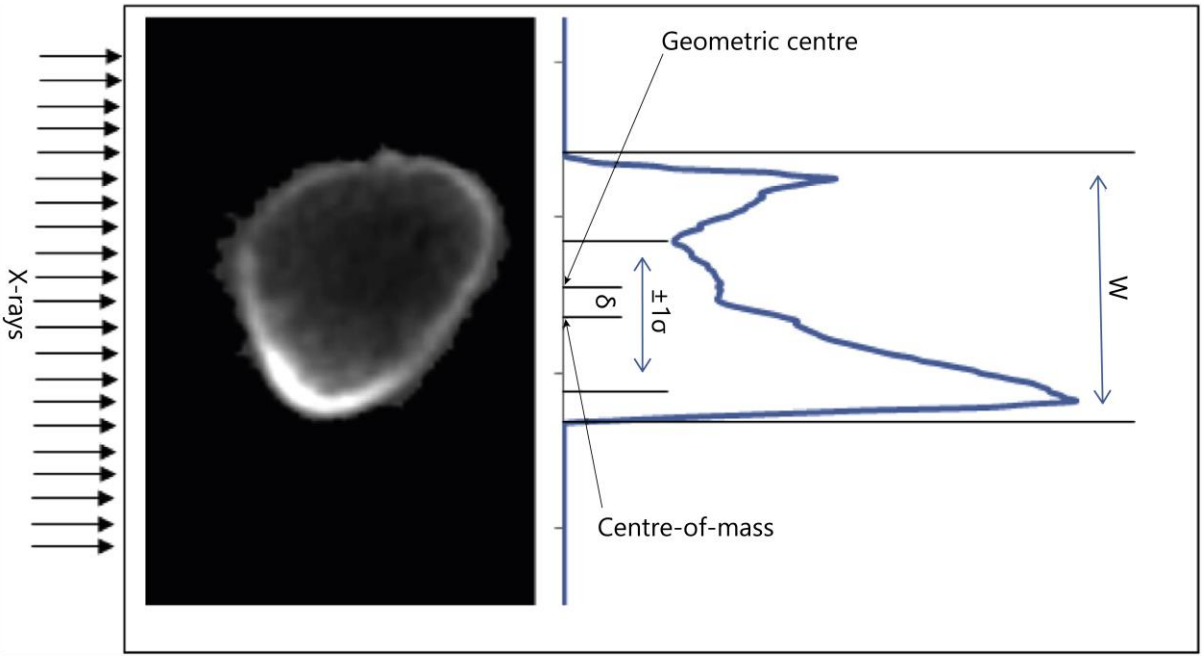


Figure 2-2. Illustration of the measures W, δ, and σ in relation to the mineral mass projection profile, where W is the periosteal width, σ is the standard deviation of a normalized mineral mass projection profile distribution, and δ the displacement between the center of mass and the geometric center of a mineral mass projection profile induced by the asymmetry of the mineral mass projection profile. Adapted from [53].

Other geometric features that are associated with hip fracture risk are hip axis length (HAL) and neck shaft angle (NSA). Both can be extracted from DXA images of the proximal femur. HAL is defined as the distance from the base of the greater trochanter to the inner pelvic rim, and NSA is the obtuse angle created by the lines of the intersection from the femoral shaft and femoral neck. In a large study including 50,420 women aged 40 and above, both HAL and NSA were robust predictors of incident hip fracture, even when corrected for age and BMD [59].

However, there are some important limitations of using DXA images for HSA. Firstly, HSA attempts to measure three-dimensional geometry in a two-dimensional projection image. This means that positioning greatly affects measured outcomes. The proximal end of the femur rotates around the acetabulum, which is not located on its long axis. Small changes in femur rotation have a large effect on projectional dimensions from which the geometry is measured. In the ideal situation, the femur is positioned so that the plane of the neck-shaft angle is parallel to the scan table. However, this is not easy to accomplish without being able to see the femur itself. Therefore, the precision of HSA is approximately 1.5 to 2 times worse than conventional BMD on the same scans [52]. The two-

dimensional nature of DXA also means that strength assessment is done only in the image plane. Since many bone cross-sections are not axially symmetrical, these values vary with rotation of the bone. DXA scans are also relatively noisy and blurred, making accurate edge detection difficult, which further decreases precision.

Finally, geometric features have to be extracted manually or semi-automatically, inducing unwanted variability, and also requiring a lot of time, making these methods impractical for large study cohorts. In an attempt to overcome these limitations, Hussain et al. developed a fully automatic method to extract hip geometry features from DXA images using a regional random forest model [60]. This model was trained to locate several anatomical landmarks, which were then used to extract 13 parameters describing the geometry of the hip, including HAL and NSA with high accuracy. This method was able to calculate the parameters with 95.87% accuracy when compared to manual extraction in the test set of 150 DXA images.

2.6. Vertebral fracture assessment

Presentation of a vertebral fracture, without major trauma or local disease, is a strong indicator for osteoporosis. A vertebral fracture is also an independent predictor of subsequent osteoporotic fractures, not only in the spine but also the hip [61, 62], and vertebral fractures are associated with a four- to fivefold increase in risk of a subsequent vertebral fracture [62 - 64]. Therefore, accurate diagnosis is of great importance for fracture prevention. Vertebral fractures are difficult to identify clinically, since symptoms are often absent or unspecific. The diagnosis for a vertebral fracture therefore relies on radiographic assessment. A vertebral fracture is indicated by an alteration in the shape and size of the vertebral body, with a reduction of height as a wedge, concave, or collapse deformity [65]. Lateral spine radiographs are the most important diagnostic tool for the detection of vertebral fractures, but in the previous decade spinal assessment using DXA equipment has gained popularity. This technique, in which a lateral image of the thoracolumbar spine is made by adapting the equipment or laterally positioning the patient, is called Vertebral Fracture Assessment (VFA).

VFA is able to support fracture risk assessment in clinical practice. Schousbou et al. performed VFA in a selected cohort of patients with BMD T-scores ≤–1.5 and age 70 and above or younger with historical height loss, or corticosteroid use [66]. In total, 9972 patients were included (93% women), of which 1575 had a vertebral fracture identified on VFA. Over a mean follow-up period of 2.8 (SD 1.7) years, 226 (2.3%) had an incident hip fracture, 715 (7.2%) had an incident non-vertebral fracture, 552 (5.5%) had a major osteoporotic fracture, and 93 (0.9%) had an incident clinical vertebral fracture. After adjustment for age, BMD, body mass index, prior fracture, parental hip fracture, glucocorticoid use, alcohol use, smoking, and rheumatoid arthritis, the presence of a vertebral fracture on VFA predicted an increased risk for incident hip fracture (HR 1.95, 95% CI 1.45–2.62), nonvertebral fracture (HR 1.99, 95% CI 1.68–2.35) and clinical vertebral fracture (HR 2.68, 95% CI 1.69–4.23).

Traditionally, images of the thoracolumbar spine were visually evaluated by clinicians to identify vertebral fractures. However, this qualitative approach is regarded as subjective and therefore may lead to disagreement, especially when performed by inexperienced observers. To overcome this problem, several standardized methods for the evaluation of vertebral fractures were developed, either based on quantitative morphometry, semi-quantitative assessment, or algorithm-based qualitative assessment. However, even fully quantitative methods require manual intervention for performing or checking morphometric measurements, and therefore still include a subjective component.

To overcome this, several fully automated methods were developed to segment vertebrae in VFA images to calculate the corresponding morphometric parameters, using active appearance models [67],

active shape models [68, 69], or random forest regression voting [70]. These methods show good segmentation accuracy for normal vertebrae, with generally deviations lower than 1 mm. Notably, both accuracy and precision are lower for fractured vertebrae, decreasing with severity grade.

Roberts et al. extracted shape and texture parameters from the outlines of vertebrae in 360 VFA images and used this to create a classification model [71]. Linear discriminants were trained to classify vertebral fractures using appearance (shape and texture) information, shape information only, and height ratios (corresponding to 6-point morphometry) only. Classification using appearance information was superior to the other methods, reaching an overall specificity of 92% at 95% sensitivity and an area under the receiver operator characteristic curve (AUC) of 0.9789. Across all fracture grades and locations, the performance of classifiers using shape parameters or height ratios reached similar performance (AUC 0.9756 for shape and 0.9651 for height ratios). Differences were larger for grade 1 fractures. For grade 2 and 3 fractures, all classifiers reached AUCs above 0.99.

Roberts et al. later combined their classification model with an automatic segmentation model to create a fully automatic computer aided diagnosis method for the detection of vertebral fractures [72]. Classification performance using automatic segmentation was compared to classification using manual segmentations. Overall the sensitivity at 95% specificity is reduced from around 95% with manual segmentation to 88% with automatic segmentation. The authors ascribe this difference in sensitivity to segmentation errors. Improving segmentation may decrease this difference in performance. Nevertheless, appearance based classification method is a promising quantitative method for the identification of vertebral fractures using VFA.

With deep learning using Convolutional Neural Networks (CNN), images can be used as inputs to a neural network, allowing for classification of images without the need for segmentation or manual feature extraction. Derkatch et al. used this technology to identify vertebral fractures by training an ensemble of CNNs on 8920 DXA-VFA images and testing it on 3822 different images [73]. This resulted in an AUC of 0.94 (95% CI 0.93 - 0.95), corresponding with a sensitivity of 87.4% and a specificity of 88.4%. In addition, adjusted hazard ratios for incident non-vertebral fracture was higher for patients with a vertebral fracture, both when detected by the CNN (1.7; 95% CI 1.3 - 2.2) and by human expert readers (1.8; 95% CI 1.3 - 2.3). When the CNN diagnosed a vertebral fracture but the expert reader did not, there was a higher risk of incident fracture but this was not statistically significant (1.3; 95% CI: 0.8 - 1.9; p=0.28), whereas no higher risk was observed when the expert reader diagnosed a vertebral fracture and the CNN did not.

A major downside of VFA is that it requires lateral imaging, and therefore additional acquisition time and radiation exposure. In an attempt to overcome this limitation, Mehta et al. used a machine learning approach to identify vertebral fractures using a standard anterior-posterior spine BMD measurement protocol [74]. Parameters extracted during lumbar spine BMD measurement include BMD, T-scores and Z-scores, and also height, width and area per vertebra from L1 to L4. These parameters, combined with hip BMD parameters and clinical characteristics (age, sex, height, weight), were used to train a Support Vector Machine (SVM) classification model. On a separate test set, the trained classification model reached an accuracy of 91.8% and an AUC of 0.8963 at a sensitivity of 81.8% and specificity of 97.4%. This method may be a promising method to screen for vertebral fractures when spinal BMD measurement is indicated, without any additional imaging.

2.7. Multivariate Fracture Risk Prediction

It is clear that the overall risk of developing osteoporotic fractures emanates from many clinical risk factors, and a single risk factor is not sufficient to identify those who will benefit from preventative

treatment. Therefore, several risk assessment tools have been developed to estimate overall fracture risk, using clinical risk factors and parameters determined using DXA imaging [14].

Currently the most widely used tool is the Fracture Risk Assessment Tool (FRAX) [75]. This tool has been incorporated into clinical practice guidelines and uses a number of risk factors to estimate the 10-year probability of developing a major osteoporotic fracture (hip, spine, forearm, or proximal humerus) and hip fracture alone. An important feature of FRAX is that it can be calibrated to different populations defined by country and ethnicity, because the risks of fracture and death vary in different regions of the world [76]. Risk factors included in FRAX are age, sex, body-mass index, previous fragility fracture, glucocorticoid use ≥3 months, secondary osteoporosis, rheumatoid arthritis, parental hip fracture, current cigarette smoking, alcohol intake of ≥3 units per day, and femoral neck BMD.

The use of BMD in FRAX is optional, allowing the tool to be used to select patients for BMD measurement as well. As such, clinical practice guidelines incorporating FRAX differ in how they use FRAX. In some guidelines, for example that of the UK National Osteoporosis Guidelines Group, FRAX is first used without BMD to estimate fracture risk. Individuals with very low fracture risk can be managed without treatment, and those with high fracture risk are considered for treatment. The intermediate group then undergoes BMD measurement followed by repeated FRAX risk assessment [77].

Other guidelines adopt a different approach, using BMD screening in the older population and considering treatment for those with BMD scores in the osteoporotic range, or those in the osteopenic range with elevated fracture risk as determined by FRAX [78]. Large randomized controlled trials have shown that screening with FRAX and initiating treatment for patients at risk of fracture can reduce the incidence of fractures [79, 80]. However, an important limitation of the current FRAX tool is that some input parameters are dichotomous and therefore do not take dose-response relations. For example, the presence of previous fracture is a binary parameter, while the number, location, and severity of previous fractures all influence the risk of subsequent fractures [81]. Also, several known independent clinical risk factors are not included in FRAX. These include past falls [82], frailty [83 - 85], and type II diabetes [86]. Calls have gone up to address these limitations in a newer version of FRAX [81].

Additionally, the only DXA derived feature included in FRAX is femoral neck BMD. As mentioned earlier, TBS, Hip Geometry features, and VFA can predict osteoporotic fractures independently from BMD and clinical parameters that are included in FRAX. Therefore, combining FRAX with one or more of these DXA derived features in a multivariate prediction model may improve its predictive value.

In a meta-analysis by McCloskey et al., combining TBS with BMD and FRAX clinical risk factors improved overall fracture prediction, especially for hip fractures [87]. For hip fractures, FRAX alone had a GR of 2.22 (95% CI 2.00–2.47). Adding TBS resulted in a slight improvement to 2.25 (95% CI 2.03 - 2.51). A similar improvement was seen for all major osteoporotic fractures, from 1.70 (95% CI 1.60 - 1.81) for FRAX alone to 1.76 (95% CI 1.65 - 1.87) for FRAX and TBS.

Similarly, in the Manitoba cohort, adding TBS to FRAX risk assessment showed improved classification of those who developed a fracture in the follow-up period [48]. The Net Reclassification Improvement, a measure of change in classification when adding an additional risk factor, was 1.2% for all major osteoporotic fractures, and 1.7% for hip fractures alone.

2.8. Artificial Intelligence for fracture prediction

In recent years, developments to fracture risk prediction have focused on the application of AI. Since AI models are good at solving nonlinear multivariate problems, AI may be particularly valuable in multifactorial osteoporotic fracture prediction.

Kruse et al. used an unsupervised machine learning technique to identify clusters of low, average and high risk of fracture for women who underwent DXA BMD measurement [88]. Spinal and hip BMD, age, body mass index, medication reimbursement, primary healthcare sector use and comorbidity of subjects were combined and clustering was done using a hierarchical agglomerative clustering algorithm. Nine clusters were identified: four clusters represented postmenopausal women with high fracture risk profiles of low BMD and between-group differences of poor versus good antiresorptive treatment compliance. One cluster of 9% was particularly worrisome due to poor treatment compliance and generalized very low BMD. Three clusters representing the majority were women with average fracture risk profiles. Two clusters of peri-menopausal and very young women represented low fracture risk subjects with high BMD and lower comorbidity.

Supervised machine learning involves training classification or regression models using a set of training data with known outcomes, either categorical (classification) or continuous (regression). Supervised classification has successfully been applied to predict fracture incidence.

Kruse et al. used data from 5439 Danish patients who underwent DXA BMD measurement and identified patients that developed a hip fracture in a 5-year follow-up period [89]. Hip and spine BMD data was paired with data on hospital admissions, medication reimbursement, comorbidities, epidemiology, and socioeconomic status. This data was used to train a variety of different classification models to predict 5-year hip fracture incidence for men and women.

Performance of classification models should be evaluated on data that is not used to train a model. Therefore, a dataset is often split in a training set and a test set. On the test set of the female cohort, bootstrap aggregated flexible discriminant analysis model reached the best performance with an AUC of 0.91 (95% CI 0.88 - 0.94). This corresponded with a sensitivity of 88% and a specificity of 81%. On the smaller male cohort, eXtreme Gradient Boosting reached a test AUC of 0.89 (95% CI 0.82 - 0.95). In this study, a 5-year follow-up was used, making it troublesome to compare its performance to the FRAX 10-year estimate. However, it is expected that this method will lead to even higher sensitivity when using 10-year follow-up, since more fractures would be included.

Ho-Le et al. evaluated an Artificial Neural Network trained on BMD data, clinical factors, and lifestyle factors to predict hip fractures in a 10-year follow-up period for 1167 post-menopausal women [90]. This strategy yielded an AUC of 0.94 on the test set, with a sensitivity of 83.3% and a specificity of 87.7%.

Both are improvements on FRAX, which has an AUC for hip fracture prediction reported in literature between 0.73 and 0.85 when including the Femoral Neck BMD score [91 - 93]. However, both of these methods only predict hip fractures, and its ability to predict other fractures is unknown.

2.9. Knowledge gaps and future research

The application of AI to fracture risk assessment with DXA has shown promising results, both for automatically analyzing images and extracting quantitative parameters, and for improving prediction algorithms to identify patients at risk of fracture.

Currently, no studies have evaluated the effect of combining the DXA derived risk factors besides BMD. Although studies have shown that VFA, HSA, and TBS are all independent of BMD, only TBS has been shown to improve fracture risk prediction when combined with BMD and clinical risk factors. It remains subject to further investigation of VFA and HSA are also of added predictive value, independently from TBS and each other.

Although still in development, integration of automatic image analysis techniques and fracture prediction algorithms could create a comprehensive workflow for fracture risk assessment without time-intensive and subjective human effort.

DXA has its place in the assessment of bone health, but the use of other imaging modalities may not be overlooked. For example, opportunistic screening for osteoporosis on Computed Tomography imaging acquired for other indications may be useful to identify patients at risk of fracture that would otherwise not have been found [94].

Also the image quality of conventional radiography remains substantially superior to DXA image quality. Recent research has shown that texture analysis of conventional radiographs could also be used as a measure of bone quality for treatment response monitoring [95]. However, further improvements in DXA image quality, combined with its very low radiation dose, might void the need for spine radiography in the future.

2.10. Conclusion

Many risk factors contribute to overall fracture risk, and besides the established risk factors such as BMD, several other risk factors can be determined using DXA. Features describing bone macrogeometry, microgeometry, and fracture status show promising results. Incorporating these risk factors into multivariate prediction models could improve the identification of those at risk of fracture and help in clinical decision making. Although still in development, AI has been successfully applied in aid of fracture risk assessment. Further research and independent external validation is needed before this can be adopted in clinical practice.

2.11. References

1. Klibanski A, Adams-Campbell L, Bassford T, Blair SN, Boden SD, Dickersin K, Gifford DR, Glasse L, Goldring SR, Hruska K, Johnson SR (2001). Osteoporosis prevention, diagnosis, and therapy. JAMA. 285(6):785-95.
2. Marshall D, Johnell O, Wedel H (1996). Meta-analysis of how well measures of bone mineral density predict occurrence of osteoporotic fractures. Bmj. 312(7041):1254-1259.
3. Wade SW, Strader C, Fitzpatrick LA, Anthony MS, O'Malley CD (2014). Estimating prevalence of osteoporosis: examples from industrialized countries. Arch Osteoporos. 9(1):182.
4. Johnell O, Kanis JA (2006). An estimate of the worldwide prevalence and disability associated with osteoporotic fractures. Osteoporos Int. 17(12):1726-1733.
5. Melton LJ, Atkinson EJ, O'Connor MK, O'Fallon WM, Riggs BL (1998). Bone density and fracture risk in men. J Bone Miner Res. 13(12):1915-1923.
6. Melton LJ, Chrischilles EA, Cooper C, Lane AW, Riggs BL (1992). Perspective: How many women have osteoporosis? J Bone Miner Res. 7:1005-1010.
7. Kanis JA, Johnell O, Oden A, Sernbo I, Redlund-Johnell I, Dawson A, De Laet C, Jonsson B (2000). Long-term risk of osteoporotic fracture in Malmö. Osteoporos Int. 11(8):669-674.
8. Empana JP, Dargent-Molina P, Bréart G (2004). Effect of hip fracture on mortality in elderly women: the EPIDOS prospective study. J Am Geriatr Soc. 52(5):685-690.
9. Gass M, Dawson-Hughes B (2006). Preventing osteoporosis-related fractures: an overview. Am J Med. 119(4):S3-S11.
10. Center JR, Nguyen TV, Schneider D, Sambrook PN, Eisman JA (1999). Mortality after all major types of osteoporotic fracture in men and women: an observational study. Lancet. 353(9156):878-882.
11. Lewiecki EM, Borges JLC (2006). Bone density testing in clinical practice. Arq Bras Endocrinol Metabol. 50(4):586-595.
12. Ross PD, Ettinger B, Davis JW, Melton LJ, Wasnich RD (1991). Evaluation of adverse health outcomes associated with vertebral fractures. Osteoporos Int. 1(3):134-140.
13. Shuhart CR, Yeap SS, Anderson PA, Jankowski LG, Lewiecki EM, Morse LR, Rosen HN, Weber DR, Zemel BS, Shepherd JA (2019). Executive summary of the 2019 ISCD position development conference on monitoring treatment, DXA cross-calibration and least significant change, spinal cord injury, peri-prosthetic and orthopedic bone health, transgender medicine, and pediatrics. J Clin Densitom. 22(4):453-471.
14. Compston JE, McClung MR, Leslie WD (2019). Osteoporosis. Lancet. 393(10169):364–376.
15. NIH Consensus Development Panel (1993). Consensus development conference: diagnosis, prophylaxis, and treatment of osteoporosis. Am J Med. 94(6):646-50.
16. Ramos RL, Armán JA, Galeano NA, Hernández AM, Gómez JG, Molinero JG (2012). Dual energy X-ray absorptimetry: fundamentals, methodology, and clinical applications. Radiología 54(5):410-423.
17. Dasher LG, Newton CD, Lenchik L (2010). Dual X-ray absorptiometry in today's clinical practice. Radiol Clin North Am. 48(3):541-560.
18. Albanese CV, Diessel E, Genant HK (2003). Clinical applications of body composition measurements using DXA. J Clin Densitom. 6(2):75-85.
19. Stone KL, Seeley DG, Lui LY, Cauley JA, Ensrud K, Browner WS, Nevitt MC, Cummings SR (2003). BMD at multiple sites and risk of fracture of multiple types: long-term results from the Study of Osteoporotic Fractures. J Bone Miner Res. 18(11):1947-1954.
20. Johnell O, Kanis JA, Oden A et al (2005). Predictive value of BMD for hip and other fractures. J Bone Miner Res. 20(7):1185-1194.

21. Dalle Carbonare L, Giannini S (2004). Bone microarchitecture as an important determinant of bone strength. J Endocrinol Invest. 27(1):99-105.
22. Cranney A, Jamal SA, Tsang JF, Josse RG, Leslie WD (2007). Low bone mineral density and fracture burden in postmenopausal women. CMAJ. 177(6):575-580.
23. Rosen EO, McNamara EA, Whittaker LG, Malabanan AO, Rosen HN (2018). Effect of positioning of the ROI on BMD of the forearm and its subregions. J Clin Densitom. 21(4):529-533.
24. Acharya S, Adsul N, Palukuri N, Acharya AS (2017). Caveats in diagnosis of osteoporosis. Indian J Med Spec 8(4):169-174.
25. Shi J, Lee S, Uyeda M et al (2016). Guidelines for dual energy X-Ray absorptiometry analysis of trabecular bone-rich regions in mice: improved precision, accuracy, and sensitivity for assessing longitudinal bone changes. Tissue Eng Part C Methods. 22(5):451-463.
26. Thurlow S, Oldroyd B, Hind K (2018). Effect of hand positioning on DXA total and regional bone and body composition parameters, precision error, and least significant change. J Clin Densitom. 21(3):375-382.
27. Peel NFA, Johnson A, Barrington NA, Smith TWD, Eastell R (1993). Impact of anomalous vertebral segmentation on measurements of bone mineral density. J Bone Miner Res. 8(6):719-723.
28. Hussain D, Han SM (2019). Computer-aided osteoporosis detection from DXA imaging. Comput Methods Programs Biomed. 173:87-107.
29. Schousboe JT, Wilson KE, Hangartner TN (2007). Detection of aortic calcification during vertebral fracture assessment (VFA) compared to digital radiography. PloS one. 2(8):e715.
30. Schousboe JT, Wilson KE, Kiel DP (2006). Detection of abdominal aortic calcification with lateral spine imaging using DXA. J Clin Densitom. 9(3):302-308.
31. Lewis JR, Schousboe JT, Lim WH, Wong G, Zhu K, Lim EM, Wilson KE, Thompson MD, Kiel DP, Prince RL (2016). Abdominal aortic calcification identified on lateral spine images from bone densitometers are a marker of generalized atherosclerosis in elderly women. Arterioscler Thromb Vasc Biol. 36(1):166-173.
32. Schousboe JT, Taylor BC, Kiel DP, Ensrud KE, Wilson KE, McCloskey EV (2008). Abdominal aortic calcification detected on lateral spine images from a bone densitometer predicts incident myocardial infarction or stroke in older women. J Bone Miner Res. 23(3):409-416.
33. Tsang JF, Leslie WD (2007). Exclusion of focal vertebral artifacts from spine bone densitometry and fracture prediction: a comparison of expert physicians, three computer algorithms, and the minimum vertebra. J Bone Miner Res. 22(6):789-798.
34. Seeman E, Delmas PD (2006). Bone quality—the material and structural basis of bone strength and fragility. N Engl J Med. 354(21):2250-2261.
35. Legrand E, Chappard D, Pascaretti C, Duquenne M, Krebs S, Rohmer V, Basle MF, Audran M (2000). Trabecular bone microarchitecture, bone mineral density, and vertebral fractures in male osteoporosis. J Bone Miner Res. 15(1):13-19.
36. Pothuaud L, Carceller P, Hans D (2008). Correlations between grey-level variations in 2D projection images (TBS) and 3D microarchitecture: applications in the study of human trabecular bone microarchitecture. Bone 42(4):775-787.
37. Silva BC, Leslie WD, Resch H, Lamy O, Lesnyak O, Binkley N, McCloskey EV, Kanis JA, Bilezikian JP (2014). Trabecular bone score: a noninvasive analytical method based upon the DXA image. J Bone Miner Res, 29(3):518-530.
38. Hans D, Barthe N, Boutroy S, Pothuaud L, Winzenrieth R, Krieg MA (2011). Correlations between trabecular bone score, measured using anteroposterior dual-energy X-ray

absorptiometry acquisition, and 3-dimensional parameters of bone microarchitecture: an experimental study on human cadaver vertebrae. J Clin Densitom. 14(3):302-312.
39. Muschitz C, Kocijan R, Haschka J, Pahr D, Kaider A, Pietschmann P, Resch H (2015). TBS reflects trabecular microarchitecture in premenopausal women and men with idiopathic osteoporosis and low-traumatic fractures. Bone 79:259-266.
40. Roux JP, Wegrzyn J, Boutroy S, Bouxsein ML, Hans D, Chapurlat R (2013). The predictive value of trabecular bone score (TBS) on whole lumbar vertebrae mechanics: an ex vivo study. Osteoporos Int. 24(9):2455-2460.
41. Maquer G, Musy SN, Wandel J, Gross T, Zysset PK (2015). Bone volume fraction and fabric anisotropy are better determinants of trabecular bone stiffness than other morphological variables. J Bone Miner Res. 30(6):1000-1008.
42. Maquer G, Lu Y, Dall'Ara E, Chevalier Y, Krause M, Yang L, Zysset PK (2016). The initial slope of the variogram, foundation of the trabecular bone score, is not or is poorly associated with vertebral strength. J Bone Miner Res. 31(2):341-346.
43. Silva BC, Walker MD, Abraham A, Boutroy S, Zhang C, McMahon DJ, Bilezikian JP (2013). Trabecular bone score is associated with volumetric bone density and microarchitecture as assessed by central QCT and HRpQCT in Chinese American and white women. J Clin Densitom. 16(4):554-561.
44. Popp AW, Buffat H, Eberli U, Lippuner K, Ernst M, Richards RG, Windolf M (2014). Microstructural parameters of bone evaluated using HR-pQCT correlate with the DXA-derived cortical index and the trabecular bone score in a cohort of randomly selected premenopausal women. PloS one. 9(2):e88946.
45. Amstrup AK, Jakobsen NFB, Moser E, Sikjaer T, Mosekilde L, Rejnmark L (2016). Association between bone indices assessed by DXA, HR-pQCT and QCT scans in post-menopausal women. J Bone Miner Metabol. 34(6):638-645.
46. Bousson V, Bergot C, Sutter B, Levitz P, Cortet B (2012). Trabecular bone score (TBS): available knowledge, clinical relevance, and future prospects. Osteoporos Int. 23(5):1489-1501.
47. Martineau P, Silva BC, Leslie WD (2017). Utility of trabecular bone score in the evaluation of osteoporosis. Curr Opin Endocrinol Diabetes Obes. 24(6):402-410.
48. Martineau P, Leslie WD, Johansson H, Harvey NC, McCloskey EV, Hans D, Kanis JA (2018). In which patients does lumbar spine trabecular bone score (TBS) have the largest effect? Bone 113:161-168.
49. McCloskey EV, Odén A, Harvey NC, Leslie WD, Hans D, Johansson H, Elders PJ (2016). A meta-analysis of trabecular bone score in fracture risk prediction and its relationship to FRAX. J Bone Miner Res. 31(5):940-948.
50. Beck TJ (2007). Extending DXA beyond bone mineral density: understanding hip structure analysis. Curr Osteoporos Rep. 5(2):49-55.
51. Beck TJ, Ruff CB, Warden KE, Scott Jr WW, Rao GU (1990). Predicting femoral neck strength from bone mineral data: a structural approach. Investigative radiology. 25(1):6-18.
52. Khoo BC, Beck TJ, Qiao QH, Parakh P, Semanick L, Prince RL, Singer KP, Price RI (2005). In vivo short-term precision of hip structure analysis variables in comparison with bone mineral density using paired dual-energy X-ray absorptiometry scans from multi-center clinical trials. Bone 37(1):112-121.
53. Khoo BC, Brown K, Zhu K, Price RI, Prince RL (2014). Effects of the assessment of 4 determinants of structural geometry on QCT-and DXA-derived hip structural analysis measurements in elderly women. J Clin Densitom. 17(1):38-46.

54. LaCroix AZ, Beck TJ, Cauley JA, Lewis CE, Bassford T, Jackson R, Chen Z (2010). Hip structural geometry and incidence of hip fracture in postmenopausal women: what does it add to conventional bone mineral density? Osteoporos Int. 21(6):919-929.
55. Kaptoge S, Beck TJ, Reeve J, Stone KL, Hillier TA, Cauley JA, Cummings SR (2008). Prediction of incident hip fracture risk by femur geometry variables measured by hip structural analysis in the study of osteoporotic fractures. J Bone Miner Res. 23(12):1892-1904.
56. Rivadeneira F, Zillikens MC, De Laet CE, Hofman A, Uitterlinden AG, Beck TJ, Pols HA (2007). Femoral neck BMD is a strong predictor of hip fracture susceptibility in elderly men and women because it detects cortical bone instability: the Rotterdam Study. J Bone Miner Res. 22(11):1781-1790.
57. Szulc P, Duboeuf F, Schott AM, Dargent-Molina P, Meunier PJ, Delmas PD (2006). Structural determinants of hip fracture in elderly women: re-analysis of the data from the EPIDOS study. Osteoporos Int. 17(2):231-236.
58. Khoo BC, Lewis JR, Brown K, Prince RL (2016). Evaluation of a simplified hip structure analysis method for the prediction of incident hip fracture events. Osteoporos Int. 27(1):241-248.
59. Leslie WD, Lix LM, Morin SN, Johansson H, Odén A, McCloskey EV, Kanis JA (2015). Hip axis length is a FRAX-and bone density-independent risk factor for hip fracture in women. The J Clin Endocrinol Metab. 100(5):2063-2070.
60. Hussain D, Han SM, Kim TS (2019). Automatic hip geometric feature extraction in DXA imaging using regional random forest. J Xray Sci Technol. 27(2):207-236.
61. Melton LJ, Atkinson EJ, Cooper C, O'Fallon WM, Riggs BL (1999). Vertebral fractures predict subsequent fractures. Osteoporos Int. 10(3):214-221.
62. Black DM, Arden NK, Palermo L, Pearson J, Cummings SR (1999). Prevalent vertebral deformities predict hip fractures and new vertebral deformities but not wrist fractures. J Bone Miner Res. 14(5):821-828.
63. Cauley JA, Hochberg MC, Lui LY, Palermo L, Ensrud KE, Hillier TA, Cummings SR (2007). Long-term risk of incident vertebral fractures. Jama. 298(23):2761-2767.
64. Lindsay R, Silverman SL, Cooper C, Hanley DA, Barton I, Broy SB, Stracke H (2001). Risk of new vertebral fracture in the year following a fracture. Jama. 285(3):320-323.
65. Guglielmi G, Diacinti D, Van Kuijk C, Aparisi F, Krestan C, Adams JE, Link TM (2008). Vertebral morphometry: current methods and recent advances. Eur Rad. 18(7):1484-1496.
66. Schousboe JT, Lix LM, Morin SN, Derkatch S, Bryanton M, Alhrbi M, Leslie WD (2019). Prevalent vertebral fracture on bone density lateral spine (VFA) images in routine clinical practice predict incident fractures. Bone 121:72-79.
67. Roberts M, Cootes TF, Adams JE (2006). Vertebral morphometry: semiautomatic determination of detailed shape from dual-energy X-ray absorptiometry images using active appearance models. Investigative radiology 41(12):849-859.
68. Van der Velde R, Ozanian T, Dumitrescu B, Haslam J, Staal J, Brett A, Geusens P. (2015). Performance of statistical models of shape and appearance for semiautomatic segmentations of spinal vertebrae T4–L4 on digitized vertebral fracture assessment images. Spine J. 15(6):1248-1254.
69. Smyth PP, Taylor CJ, Adams JE (1999). Vertebral shape: automatic measurement with active shape models. Radiology 211(2):571-578.
70. Bromiley PA, Adams JE, Cootes TF (2015). Automatic localisation of vertebrae in DXA images using random forest regression voting. International Workshop and Challenge on Computational Methods and Clinical Applications for Spine Imaging 9402:38-51
71. Roberts M, Cootes T, Pacheco E, Adams J (2007). Quantitative vertebral fracture detection on DXA images using shape and appearance models. Academic radiology. 14(10):1166-1178.

72. Roberts MG, Pacheco EMB, Mohankumar R, Cootes TF, Adams JE (2010). Detection of vertebral fractures in DXA VFA images using statistical models of appearance and a semi-automatic segmentation. Osteoporos Int. 21(12):2037-2046.
73. Derkatch S, Kirby C, Kimelman D, Jozani MJ, Davidson JM, Leslie WD (2019). Identification of vertebral fractures by convolutional neural networks to predict nonvertebral and hip fractures: a registry-based cohort study of dual X-ray absorptiometry. Radiology 293(2):405-411.
74. Mehta SD, Sebro R (2020). Computer-aided detection of incidental lumbar spine fractures from routine dual-energy X-ray absorptiometry (DXA) studies using a support vector machine (SVM) classifier. J Digit Imaging. 33(1):204-210.
75. Kanis JA, McCloskey EV, Johansson H, Oden A, Ström O, Borgström F (2010). Development and use of FRAX® in osteoporosis. Osteoporos Int. 21(2):407-413.
76. Kanis JA, Johnell O, De Laet C, Jonsson B, Oden A, Ogelsby AK (2002). International variations in hip fracture probabilities: implications for risk assessment. J Bone Miner Res. 17(7):1237-1244.
77. Compston J, Cooper A, Cooper C, Gittoes N, Gregson C, Harvey N, Reid DM (2017). UK clinical guideline for the prevention and treatment of osteoporosis. Arch Osteoporos. 12(1):43.
78. Cosman F, de Beur SJ, LeBoff MS, Lewiecki EM, Tanner B, Randall S, Lindsay R. (2014). Clinician's guide to prevention and treatment of osteoporosis. Osteoporos Int. 25(10):2359-2381.
79. Rubin KH, Rothmann MJ, Holmberg T, Høiberg M, Möller S, Barkmann R, Brixen K (2018). Effectiveness of a two-step population-based osteoporosis screening program using FRAX: the randomized Risk-stratified Osteoporosis Strategy Evaluation (ROSE) study. Osteoporos Int. 29(3):567-578.
80. Shepstone L, Lenaghan E, Cooper C, Clarke S, Fong-Soe-Khioe R, Fordham R, Holland R (2018). Screening in the community to reduce fractures in older women (SCOOP): a randomised controlled trial. Lancet. 391(10122):741-747.
81. El Miedany Y (2020). FRAX: re-adjust or re-think. Arch Osteoporos. 15(1):1-8.
82. Harvey NC, Odén A, Orwoll E, Lapidus J, Kwok T, Karlsson MK, Kanis JA (2018). Falls predict fractures independently of FRAX probability: a meta-analysis of the Osteoporotic Fractures in Men (MrOS) Study. J Bone Miner Res. 33(3):510-516.
83. Ensrud KE, Ewing SK, Taylor BC, Fink HA, Cawthon PM, Stone KL, Tracy JK (2008). Comparison of 2 frailty indexes for prediction of falls, disability, fractures, and death in older women. Arch Internal Med. 168(4):382-389.
84. Kennedy CC, Ioannidis G, Rockwood K, Thabane L, Adachi JD, Kirkland S, Papaioannou A (2014). A Frailty Index predicts 10-year fracture risk in adults age 25 years and older: results from the Canadian Multicentre Osteoporosis Study (CaMos). Osteoporos Int. 25(12): 2825-2832.
85. Li G, Papaioannou A, Thabane L, Cheng J, Adachi JD (2016). Frailty Change and Major Osteoporotic Fracture in the Elderly: Data from the Global Longitudinal Study of Osteoporosis in Women 3-Year Hamilton Cohort. J Bone Miner Res. 31(4):718-724.
86. Leslie WD, Johansson H, McCloskey EV, Harvey NC, Kanis JA, Hans D (2018). Comparison of methods for improving fracture risk assessment in diabetes: the Manitoba BMD Registry. J Bone Miner Res. 33(11):1923-1930.
87. McCloskey EV, Odén A, Harvey NC, Leslie WD, Hans D, Johansson H, Elders PJ (2016). A meta-analysis of trabecular bone score in fracture risk prediction and its relationship to FRAX. J Bone Miner Res. 31(5):940-948.

88. Kruse C, Eiken P, Vestergaard P (2017). Clinical fracture risk evaluated by hierarchical agglomerative clustering. Osteoporos Int. 28(3):819-832.
89. Kruse C, Eiken P, Vestergaard P (2017). Machine learning principles can improve hip fracture prediction. Calcif Tissue Int. 100(4):348-360.
90. Ho-Le TP, Center JR, Eisman JA, Nguyen TV, Nguyen HT (2017). Prediction of hip fracture in post-menopausal women using artificial neural network approach. In 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) 4207-4210.
91. Premaor M, Parker RA, Cummings S, Ensrud K, Cauley JA, Lui LY (2013). Predictive value of FRAX for fracture in obese older women. J Bone Miner Res. 28(1):188-195.
92. Azagra R, Roca G, Encabo et al (2012). FRAX® tool, the WHO algorithm to predict osteoporotic fractures: the first analysis of its discriminative and predictive ability in the Spanish FRIDEX cohort. BMC musculoskelet Disord. 13(1):204.
93. Kälvesten J, Lui LY, Brismar T, Cummings S (2016). Digital X-ray radiogrammetry in the study of osteoporotic fractures: Comparison to dual energy X-ray absorptiometry and FRAX. Bone 86:30-35.
94. Gausden EB, Nwachukwu BU, Schreiber JJ, Lorich DG, Lane JM (2017). Opportunistic use of CT imaging for osteoporosis screening and bone density assessment: a qualitative systematic review. Jbjs. 99(18):1580-1590.
95. Dimai HP, Ljuhar R, Ljuhar D, Norman B, Nehrer S, Kurth A, Fahrleitner-Pammer A (2019). Assessing the effects of long-term osteoporosis treatment by using conventional spine radiographs: results from a pilot study in a sub-cohort of a large randomized controlled trial. Skelet Radiol. 48(7):1023-1032.

3. Inter-operator agreement of vertebral fracture assessment with dual-energy x-ray absorptiometry equipment

In collaboration with S.R. Romeijn, P. Dibbets-Schneider, D.D.D. Rietbergen, L.M. Pereira Arias-Bouda, C. Götz, M.D. DiFranco, H.P. Dimai, and W. Grootjans

3.1. Abstract

Purpose: To investigate the time and effort needed to perform vertebral morphometry, as well as interobserver agreement for identification of vertebral fractures on Vertebral Fracture Assessment (VFA) images, and to evaluate the potential benefit for automated VFA.

Methods: Ninety-six VFA images were retrospectively selected, representing a wide range of age groups and vertebral fractures of different types and severities. Three clinical radiographers independently performed semi-automatic 6-point morphometry on all images and fractures were identified and graded according to the Genant classification. The time needed to annotate each image was recorded and reader fatigue was assessed using a modified Simulator Sickness Questionnaire (SSQ). Inter-observer agreement was assessed on per-patient level and on per-vertebra level for detecting fractures of all grades (grade 1 - 3) and for moderate and severe (grade 2 and 3) fractures only using the kappa statistic. Variability in vertebral height measurement was evaluated using the intraclass correlation coefficient (ICC).

Results: Per-patient agreement was 0.59 for grade 1 - 3 fracture detection, and 0.65 when including only grade 2 -3 fractures. Agreement between radiographers for per-vertebra fracture severity classification was 0.92. Vertebral height measurements showed a mean absolute difference from the average across radiographers of 1.38 mm (95% CI: 1.36 - 1.41), with an intraclass correlation of 0.96. The time needed to annotate the VFA images ranged between 91 and 540 seconds, with a mean annotation time of 259 seconds. Mean SSQ scores were significantly lower at the start of a reading session (1.29; 95% CI: 0.81 - 1.77) compared to the end of a reading session (3.25; 95% CI: 2.60 - 3.90; $p < 0.001$).

Conclusion: Although trained radiographers performing vertebral morphometry on VFA achieve excellent intraclass correlation for vertebral height measurement, agreement for detection of patients with vertebral fractures is only moderate. This suggests that small variations in landmark placement can lead to different classifications. In addition, vertebral morphometry requires substantial time investment and significantly affects reader fatigue. There is a potential benefit for automation tools for detection of vertebral fractures on VFA, both in improving interobserver agreement and in decreasing reading time and burden on readers.

3.2. Introduction

With a current estimate of 200 million people worldwide, osteoporosis is the most common metabolic bone disease [1]. The prevalence of osteoporosis is higher in women and increases with age, from 19% among women aged 65 to 74 years to >50% in women aged ≥85 years [2-4]. With age the predominant factor associated with osteoporosis, the number of osteoporosis patients is expected to increase dramatically with the aging population [5]. Osteoporosis is defined as 'a systemic skeletal disorder characterized by a low bone mass and by microarchitectural deterioration of bone tissue, with a subsequent increase in bone fragility and susceptibility to fracture'. Vertebral fractures are the most

common osteoporotic fractures, and have a major impact on patients' quality of life by causing back-pain, reduced physical capability, poor perceived general health, and emotional status [6].

Presentation of a vertebral fracture, without major trauma or local disease, is a strong indicator for osteoporosis and an independent predictor of subsequent osteoporotic fractures, not only in the spine but also the hip [7, 8]. Corrected for age and Bone Mineral Density (BMD), a vertebral fracture is associated with a four- to fivefold increase in risk of a subsequent vertebral fracture [8-10] or hip fracture [11]. Assessment of vertebral fractures is therefore considered fundamental in management and treatment of osteoporosis and the prevention of subsequent osteoporotic fractures [12, 13].

Although conventional lateral radiography of the spine remains the gold standard for identification of vertebral fractures, densitometric vertebral fracture assessment (VFA) has some important advantages. Dual-Energy X-ray Absorptiometry (DXA) equipment is used to make a lateral spine scan, requiring very little radiation exposure (3–40 vs. 600–1600 microSieverts for spinal radiography) [14]. Specialized software allows for quantitative vertebral morphometry to identify vertebral fractures on these images. Even though VFA image resolution is lower than that of spinal radiographs, VFA has shown good sensitivity and specificity for the detection of vertebral fractures [15], and more patients with asymptomatic vertebral fractures can be identified if VFA is used systematically at the time of bone mineral density measurement. However, vertebral morphometry requires manual or semi-automatic characterization of vertebral height, which is labor intensive and may be subject to inter-observer variability, limiting widespread adoption in clinical practice. Automation of vertebral fracture detection may help overcome this problem. Question remains whether there is a business case for VFA automation tools. Therefore, this reader study investigates the time and effort needed to manually perform vertebral morphometry, as well as inter operator variability for identification of vertebral fractures on VFA images.

3.3. Methods

3.3.1. Patients

For this study, a retrospective search was conducted in the digital picture archiving and communication system (PACS) of patients referred for DXA imaging between 01-07-2019 and 31-03-2020 who underwent VFA imaging. The study protocol was approved by the Medical Ethics Committee of the Leiden University Medical Center (registration number G20.032) and the requirement to obtain written consent was waived. Of the patients who underwent VFA in this timeframe, a clinically representable group of 96 patients was selected. The group consisted of patients with a wide age range and included both patients without and with vertebral fractures of different types and severities on different vertebral levels. Indications for VFA include clinically suspected or diagnosed osteoporosis, chronic glucocorticoid therapy, and follow-up after organ transplant [16]. In our Fracture Liaison Service (FLS), patients often undergo both VFA and conventional spinal radiography on the same day. When available, these spinal radiographs were also included in the analysis.

3.3.2. Image acquisition

Postero-anterior and lateral VFA images of the thoracolumbar spine (T4 - L4) were made by trained radiographers using Hologic Horizon A DXA equipment (Hologic, Bedford, MA, USA. Software version 13.6). Patients were positioned in supine position with a cushion supporting the knees, and the c-arm was rotated for lateral imaging. Routine BMD measurements at the level of the lumbar spine (L1 - L4) and the hip were made by dual-energy absorptiometry in the same session on the same equipment. T-scores for adults and Z-scores for children were calculated from hip and spine BMD using NHANES-III reference values. The diagnosis for osteoporosis, osteopenia or normal BMD was established using the

World Health Organization criteria, with osteopenia diagnosed for T-scores or Z-scores between -1 and -2.5, and osteoporosis for T-scores or Z-scores equal to or below -2.5.

Lateral radiographs of the thoracic and lumbar spine were acquired using a standardized protocol, with the Canon CXDI detector (Canon Inc., Ota, Japan) centralized on T7 for the thoracic spine and on L3 for the lumbar spine.

3.3.3. Assessment of vertebral fractures

VFA images were independently analysed by three clinical radiographers using Hologic Physician Viewer software (Version 7.3, Hologic, Bedford, MA, USA). Radiographers had different levels of experience with VFA (20, 10 and 5 years) and were blinded to each other's assessments. The study workflow is schematically depicted in Figure 3-1. Each reader performed vertebral semi-automatic 6-point morphometry, in which the software automatically places points on the four corners and in the middle of both the upper and lower endplate of each vertebra from T4 to L4. The positions of the points are then manually adjusted by the readers. The software calculates the anterior, medial, and posterior vertebral heights and uses these to determine height ratios. The wedge ratio is calculated by dividing anterior height by posterior height, biconcavity is calculated by dividing medial height by posterior height, and the crush ratio is determined by dividing posterior height by posterior heights of adjacent vertebrae. If adjacent vertebrae are fractured, the height of the closest non-fractured vertebra is used.

After determining the height ratios, a grade (1-3) is assigned to the fracture, as defined by Genant et al., to quantify the severity of the vertebral fracture. [17]. Grade 0 (normal) is assigned for a height loss of less than 20%, grade 1 for height loss between 20% and 25% (mild), grade 2 for a height loss between 25% and 40% (moderate), and grade 3 for a height loss of more than 40% (severe). During the VFA, vertebrae that the radiographers deemed not evaluable due to insufficient image quality or image artefacts were classified as not evaluable.

Both the classification per vertebra and the coordinates of the 6 points describing its morphology were exported for analysis. To allow for accurate comparison of annotations made by the different radiographers, the images were cross-checked for differences in identification of vertebral levels. In case of discrepancies in the assigned vertebral level, corrections were made in vertebral levels where the majority vote by two of the three readers was assumed to be correct.

Conventional spinal radiographs were visually evaluated by an experienced radiologist using the Genant Semi-quantitative classification method.

3.3.4. Reader fatigue

All VFA images were presented to each radiographer in the same order, and were sequentially annotated in sessions where 6 images were annotated. Between different annotation sessions, a break of at least 15 minutes was planned. At the start and end of each session, the readers filled in a modified Simulator Sickness Questionnaire (SSQ) to assess reader fatigue [18]. The readers were asked to score the presence of 7 common symptoms regarding fatigue and eyestrain on a 5-point scale; the overall SSQ score was given as the sum of these scores. The time needed to annotate each image was also recorded.

3.3.5. Analyses and statistics

Inter-observer variability of classification of vertebral deformities using 6-point morphometry on VFA images was assessed on per-patient level and on per-vertebra level. For the per-patient analysis, each image was classified as either fractured or not fractured based on the vertebra with the highest Genant

grade in the evaluated image. As a measure of interobserver agreement, the kappa statistic was calculated for detecting fractures of all grades (grade 1 - 3) and for moderate and severe (grade 2 and 3) only. Since raters were not forced to assign a fixed number of cases to each category, Randolph's free-marginal multirater kappa was used [19]. Randolph's kappa was also calculated to determine per-vertebra agreement of fracture severity classifications. Vertebrae that one or more radiographers deemed not evaluable were discarded from the analysis.

In addition, intra-reader variability with respect to landmark placement was evaluated. This was done by comparing the absolute landmark coordinates and absolute vertebral height measurements across readers. Variability in landmark placement was expressed as Euclidean distance to the average landmark location across radiographers. Distance to the average landmark location for patients where all radiographers agreed on fracture status and for patients where there was disagreement was compared using the student's t-test. Reliability in vertebral height measurement was expressed as

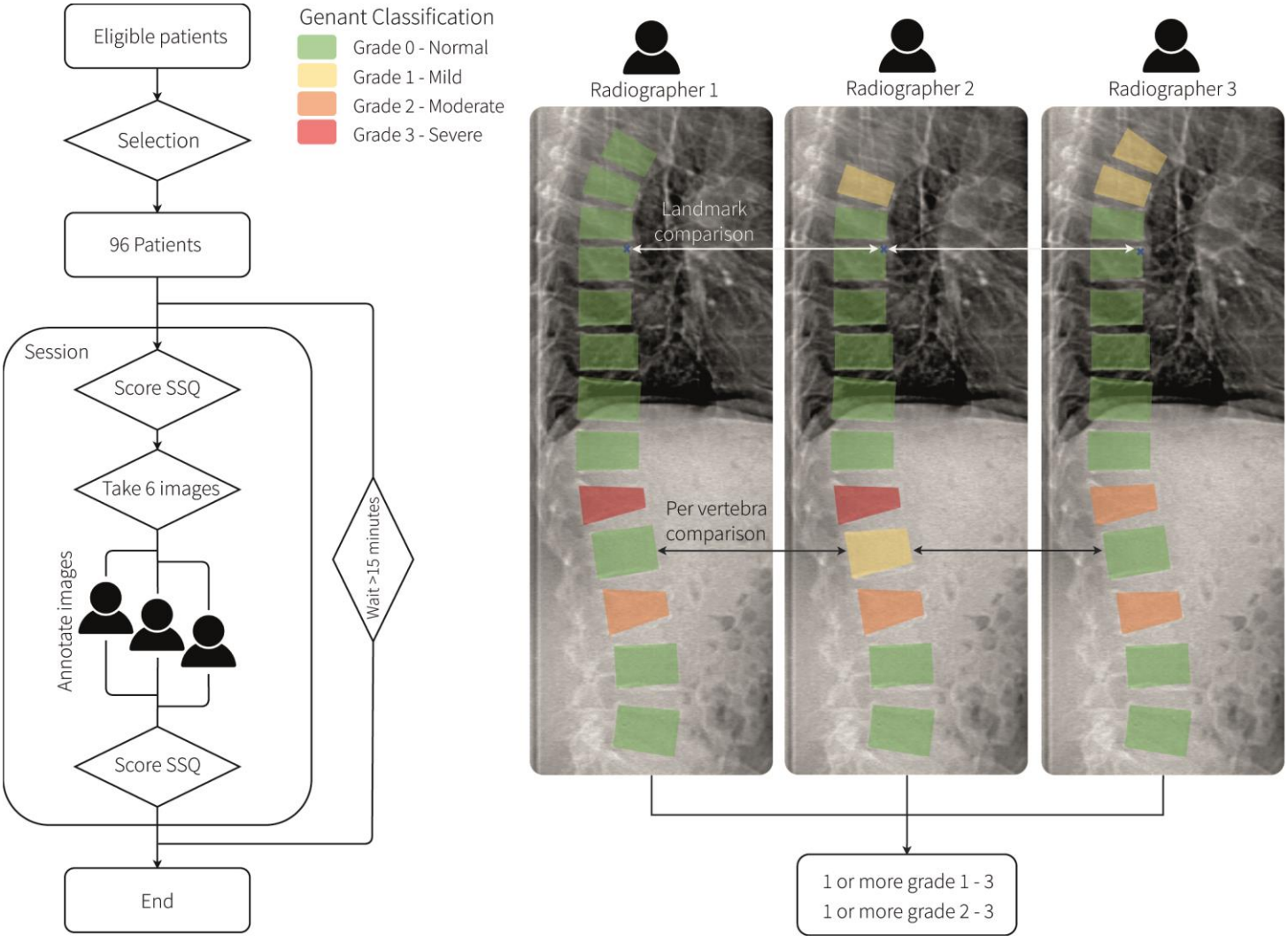


Figure 3-1. Left: Schematic description of the VFA annotation workflow in our study. Three radiographers annotated the same 96 images in sessions of 6 images, with at least 15 minutes between sessions. Reader fatigue was assessed by SSQ at the start and end of each annotation session. Annotation consisted of 6-point morphometry by semi-automatic landmark placement on vertebrae from T4 to L4, after which Genant classifications were automatically calculated. Right: annotations made by the three radiographers of the same VFA image. In color, the Genant classification (green: normal, yellow: mild, orange: moderate and red: severe). SSQ: Simulator Sickness Questionnaire. VFA: Vertebral Fracture Assessment.

intraclass correlation (ICC; fixed raters, single rating). ICC values less than 0.5 were considered indicators of poor reliability; values between 0.5 and 0.75 indicate moderate reliability, values between 0.75 and 0.9 indicate good reliability, and values greater than 0.90 indicate excellent reliability [20].

Spinal radiographs (if available) were used as a gold standard to determine the presence of a vertebral fracture. This was used to compute Cohen's kappa, sensitivity, and specificity for each radiographer individually for detecting a vertebral fracture regardless of fracture grade and for grade 2 and 3 fractures only.

For comparing the annotation efforts, the SSQ scores at the start and end of the annotation sessions were evaluated by comparing medians using a Wilcoxon Signed Rank test. Significance levels for all statistical tests were set to 0.05.

3.4. Results

VFA images of 2468 patients were made within the defined time frame. Of these, 96 images were selected and annotated by the three radiographers. Annotations for one VFA image were lost due to a data transfer error, so 95 annotated VFA images were included for analysis. Sixty-five (68%) of the included patients were female, and 30 were male. The mean age was 61.4 ± 16.0 years (range 12 - 85). Forty-four (46%) of patients had BMD in the osteopenic range, 32 (34%) had osteoporosis, and 19 (20%) had normal BMD. Discrepancies in identification of vertebral levels across radiographers were found in 16 patients (153 vertebrae), all of which could be resolved by applying the vertebral label given by the majority of the radiographers.

3.4.1. Fracture classification

3.4.1.1. Per patient analysis

All three radiographers agreed for 42 patients that no vertebral fractures were visible, and 24 patients had one or more fractures of grade 1 - 3, as agreed on by all three radiographers. In the remaining 29 patients, there was disagreement about the presence of fractures; 11 had one or more fractures detected by two radiographers, and 18 patients had a fracture detected by only one of the radiographers. This resulted in a Randoph's kappa score of 0.59. When including only grade 2 -3 fractures, 56 patients were considered not fractured, 14 were considered fractured, and there was disagreement in 25 patients, resulting in a kappa score of 0.65. Fracture classification and reader agreement are presented in Figure 3-2. Agreement between sets of two radiographers ranged between 0.53 and 0.62 (Cohen's kappa), and was higher between the two least experienced radiographers (0.62) than between the most experienced radiographer and the less experienced radiographers (0.57 & 0.53).

For 57 patients a spinal radiograph was available. Of these, 28 patients (49%) had at least one vertebral fracture grade 1 - 3, and 16 patients (28%) had one or more vertebral fractures of grade 2 - 3. Only 15 (54%) patients with grade 1 - 3 fractures and 5 (31%) patients with grade 2 - 3 fractures were detected by all three radiographers on VFA. Agreement between VFA and spinal radiography ranged between 0.51 and 0.58 for fractures regardless of severity, and between 0.52 and 0.60 when including only grade 2 - 3 fractures. When considering fracture detection on spinal radiographs as the ground truth, the radiographers detected vertebral fractures on VFA with a sensitivity ranging between 0.69 and 0.75 and a specificity ranging between 0.81 and 0.90 for detection of grade 2-3 vertebral fractures.

3.4.1.2. Per vertebra analysis

Of the 1248 vertebrae included, 121 (9.7%) vertebrae in 45 patients were considered not evaluable by one or more radiographers and were excluded for per-vertebra analyses. T4 was not evaluated most often (40), followed by T5 (18), T6 (18), T7 (15), T8 (10), L4 (8), T9 (7), T10 (3), T11 (1), and T12 (1). Randolph's kappa for agreement between radiographers for per-vertebra fracture severity classification was 0.92. When split per vertebral level, agreement was highest for L4 (0.96) and lowest for T7 (0.84). Agreement per vertebral level is shown in Figure 3-3. Agreement for T4 could not be determined since all included T4 vertebrae were considered normal by all three radiographers. Of the 121 vertebrae that were considered not evaluable by one or more radiographers, 12 vertebrae in 9 patients were classified as fractured (grade 2 - 3) by another radiographer. For 6 patients this affected the highest-grade vertebra, and would have affected fracture diagnosis.

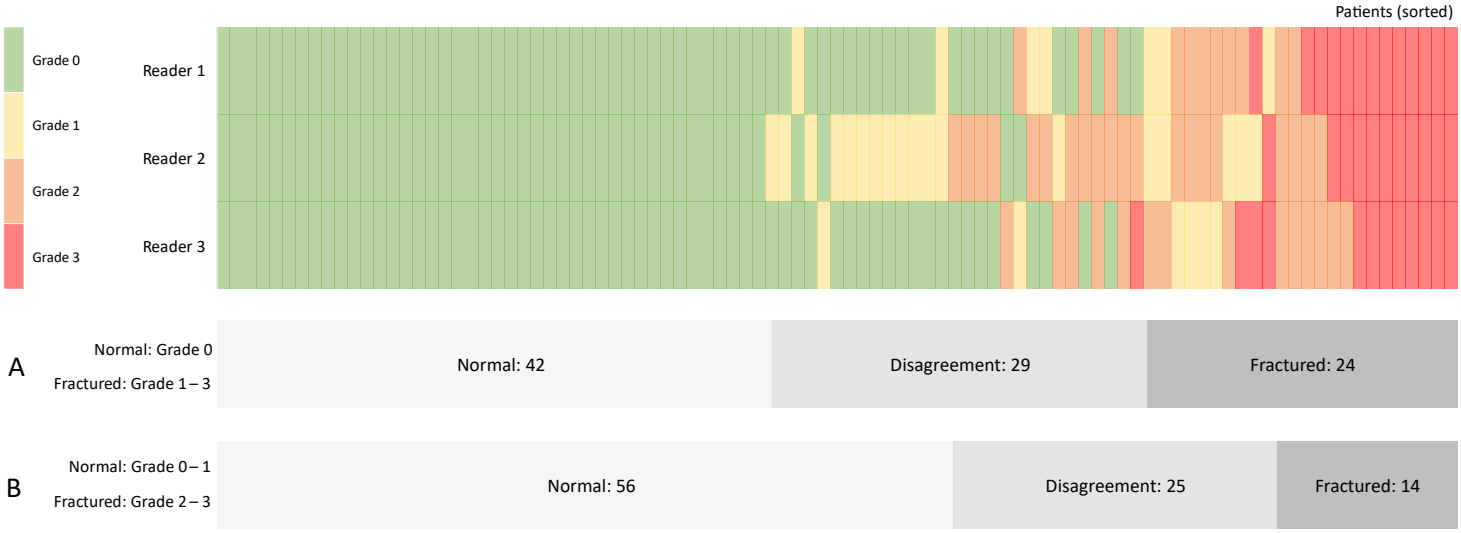


Figure 3-2. Highest detected Genant classification per patient on VFA as determined by the three radiographers. Patients are classified as fractured or not fractured, either including grade 1 (A) or excluding grade 1 (B).

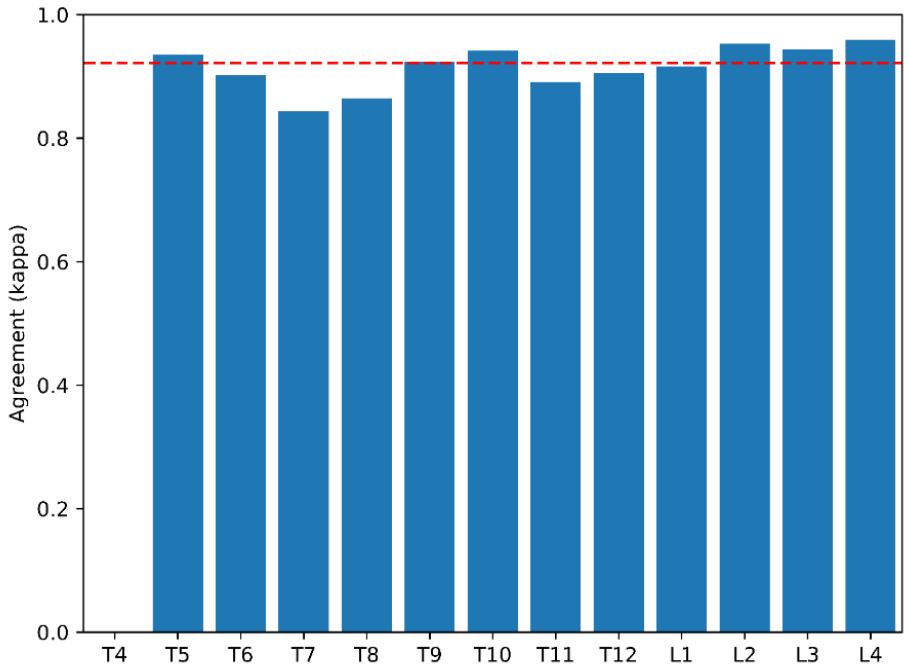


Figure 3-3. Agreement (Randolph's kappa) per vertebral level for vertebral fracture severity classification on VFA. The red dashed line indicates overall agreement level. Agreement for T4 could not be determined since all included T4 vertebrae were considered grade 0 by all radiographers.

3.4.2. Height measurement & landmark placement

Landmark placement was analyzed for 1127 vertebrae. With 6 landmarks per vertebra, this resulted in a total of 6762 sets of landmark coordinates, each set consisting one coordinate pair for each radiographer. For each landmark, the average location across radiographers was determined. The distance to the average location ranged between 0.0 and 8.36 mm, with a mean absolute distance of 0.81 mm (95% CI: 0.80 - 0.82). The spread around the mean for each landmark is shown in Figure 3-4. Spread is slightly elongated in the anteroposterior direction (x-direction) with a standard deviation of 0.84 mm compared to 0.52 mm in the craniocaudal y-direction. For patients where all radiographers agreed on the classification in fractured (grade 2-3) or not fractured (grade 0-1), the mean absolute distance was 0.79 mm (95% CI: 0.78 - 0.80). For patients where there was disagreement, this was 0.87 mm (95% CI: 0.86 - 0.89; $p < 0.001$).

Vertebral height measurements showed a mean absolute difference from the average across radiographers of 1.38 mm (95% CI: 1.36 - 1.41), with an intraclass correlation of 0.96.

3.4.3. Reading time & reader discomfort

The time needed to annotate the VFA images ranged between 91 and 540 seconds, with a mean annotation time of 259 seconds. Mean SSQ scores were significantly lower at the start of a reading session (1.29; 95% CI: 0.81 - 1.77) compared to the end of a reading session (3.25; 95% CI: 2.60 - 3.90; $p < 0.001$). SSQ scores per reader for each reading session are shown in Figure 3-5.

3.5. Discussion

3.5.1. Interpretation of results

In this study, inter-observer agreement of VFA for diagnosis of vertebral fractures was evaluated. On a per-patient level, agreement between readers was moderate for grade 2 - 3 fractures. However, when evaluating classifications on a per-vertebra level, agreement was much higher. This apparent difference can be explained by the way per-patient classifications are determined. Readers can agree that a patient has twelve non-fractured vertebrae, but disagree whether the remaining vertebra is fractured or not. In the per-vertebra analysis, agreement on 12 out of 13 vertebrae leads to a relatively high kappa score, whilst there is disagreement on the fracture status of this patient, potentially affecting clinical decision making. Another factor contributing to high per-vertebra agreement is the exclusion of vertebrae that were not evaluated by at least one reader. Not evaluating a fractured vertebra due to insufficient visibility can lead to that patient not being classified as fractured, which has happened in 6 cases in this study. Agreement was slightly lower when grade 1 fractures were also included, which seems to be induced by readers' difficulty in differentiating grade 1 deformities from normal vertebrae.

These findings provide a basis for clinically relevant performance targets for automated VFA. Landmark placement is important for accurate fracture detection and to be clinically usable, automation tools would need to be able to place landmarks with very high accuracy. However, merely placing landmarks close to where humans would place them is not sufficient, and fracture detection on a patient level should be the primary outcome measure, as this is the most critical measure of performance and has direct consequences on clinical treatment decisions.

Annotation of images for VFA purposes requires significant time and user effort. With VFA having an increasing importance in clinical decision making of patients with osteoporosis, the required effort for annotating these images can become problematic. Indeed, this study showed that radiographers require substantial time to annotate these images. Furthermore, reader fatigue significantly increased during reading sessions. Readers mainly reported eye strain symptoms, which is not unexpected given the nature of the annotation task.

3.5.2. Comparison to literature

Few studies evaluated inter-observer agreement of vertebral morphometry for the detection of vertebral fractures on VFA. Pearson et al. compared vertebral morphometry variation between two DXA systems and found good agreement between two observers in identifying severe fractures, but a lack of agreement for identifying moderate fractures [22]. Kappa scores were 0.51 and 0.79 for the two DXA systems respectively. However, this study included only 25 patients. In a similar study, Bazzochi et al. evaluated a semiquantitative method supplemented by vertebral morphometry for suspected vertebral fractures. Inter-observer agreement between the three readers ranged from 0.665 to 0.713 [23]. Dort et al. Evaluated vertebral height measurement on DXA images and found excellent ICC (>0.95) and moderate agreement for detecting vertebral fractures, with a kappa score of 0.628 for grade 2 – 3 fractures and 0.699 for grade 1 – 3 fractures [24]. Inter-observer agreement of visual semi-quantitative identification of vertebral fractures, without quantitative morphometry, has been reported between 0.51 and 0.69 [25-27]. Although kappa scores are not easily compared to other studies due to differences in patient populations, number of participants, imaging protocols, equipment, and fracture identification methods, inter-observer agreement of vertebral morphometry in this study seems to be in concordance with previously reported results.

Agreement with conventional spinal radiography, sensitivity, and specificity of VFA has been extensively reported in literature. A recent meta-analysis found a pooled sensitivity of 0.84 and a specificity of 0.90 [15]. In our study we found a sensitivity and specificity similar to that reported in literature, albeit at the lower end of the range [28].

Bazzochi et al. also reported an average VFA reader time of 23.1±16.2 seconds per vertebra. This is very similar to our findings, as an average annotation time of 259 seconds per patient gives 22.1 seconds per vertebra, when accounting for non-evaluable vertebrae.

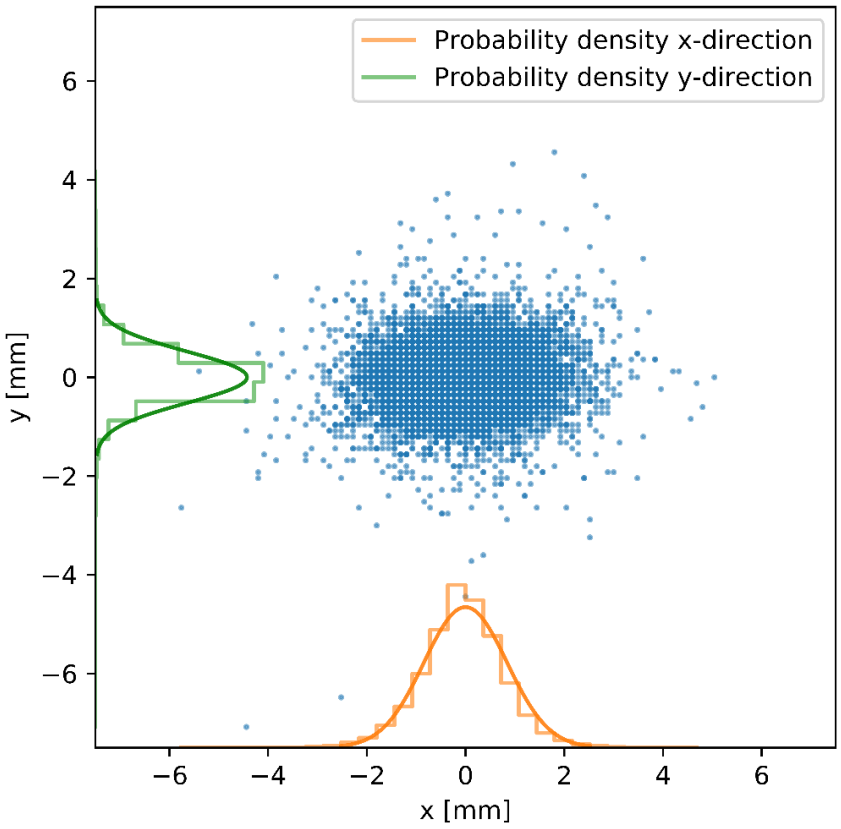


Figure 3-4. Spread around the average location for each landmark. Landmarks are translated so that the average location is at the origin, and scatter points correspond to individual annotations by one of the radiographers.

3.5.3. Business case

Detection of vertebral fractures with VFA has some important benefits compared to conventional radiography. The first advantage is a lower patient burden since scans are made in the same session using a low radiation dose. However, VFA is subject to significant inter-observer variability, and conventional spinal radiography remains the gold standard. Therefore, we believe the main application of VFA is as an additional screening tool for patients undergoing BMD measurement. It is estimated that as much as 70% of all vertebral fractures go undiagnosed [29, 30], and screening for vertebral fractures has been shown to be cost-effective [31]. Nevertheless, many patients still undergo BMD measurement without VFA. The significant time investment needed to annotate VFAs likely contributes to this, as it would currently take too much time to do VFA for all patients undergoing BMD measurement. A potential method to help solve this problem is the automation of vertebral fracture detection, allowing much more VFAs to be done without significant investments, and potentially diagnosing many vertebral fractures that would otherwise have been missed.

3.5.4. Limitations

A major limitation of our study is the fact that conventional radiographs were not available for all patients. Only 57 out of 95 patients underwent conventional radiography besides VFA. Since conventional spinal radiographs are considered the gold standard for vertebral fracture detection, this meant that patients' true fracture status was only available for a subset of patients. Most patients in this subset were treated in our FLS, since it is standard procedure in our FLS to do both VFA and spinal radiography. Patients are admitted to FLS programs after already sustaining a fracture, and these patients therefore have a higher prevalence of vertebral fractures than other populations. Our findings of moderate agreement of VFA with spinal radiography are specifically applicable to FLS patient populations.

Another limitation is the lack of a universally accepted definition of which vertebral deformity is a vertebral fracture. Every vertebral fracture is a vertebral deformity, but not every vertebral deformity is a vertebral fracture. With vertebral morphometry, qualitative features of morphology are not taken into account, and therefore in this study we measured vertebral deformities rather than vertebral fractures exclusively.

In this study, readers evaluated six VFA images in a row, and then rested at least 15 minutes before starting a new reading session. Besides these requirements, radiographers were free to choose their exact reading schedule, and could choose to do multiple sessions on the same day or spread them out across a longer period of time. From the results as shown in figure 5, it seems that baseline SSQ scores increase for a number of consecutive sessions. This cumulative fatigue may indicate that 15 minutes of rest is not enough, and longer periods between sessions would be required to get fatigue levels back to baseline. However, in this study we looked at the difference between session start and end, mitigating cumulative effects. In addition, radiographers were asked to evaluate VFA images alone and were not allowed to assist each other, which may not be representative of clinical practice.

3.5.5. Conclusion

Although multiple trained radiographers performing vertebral morphometry on VFA achieve very small differences in landmark placement and excellent intraclass correlation for vertebral height measurement, agreement for detection of patients with vertebral fractures is only moderate. This suggests that small variations in landmark placement can lead to different classifications. There is a potential benefit for automation tools for detection of vertebral fractures on VFA, but automation tools should focus on clinically relevant outcome measures such as agreement with conventional radiographic imaging.

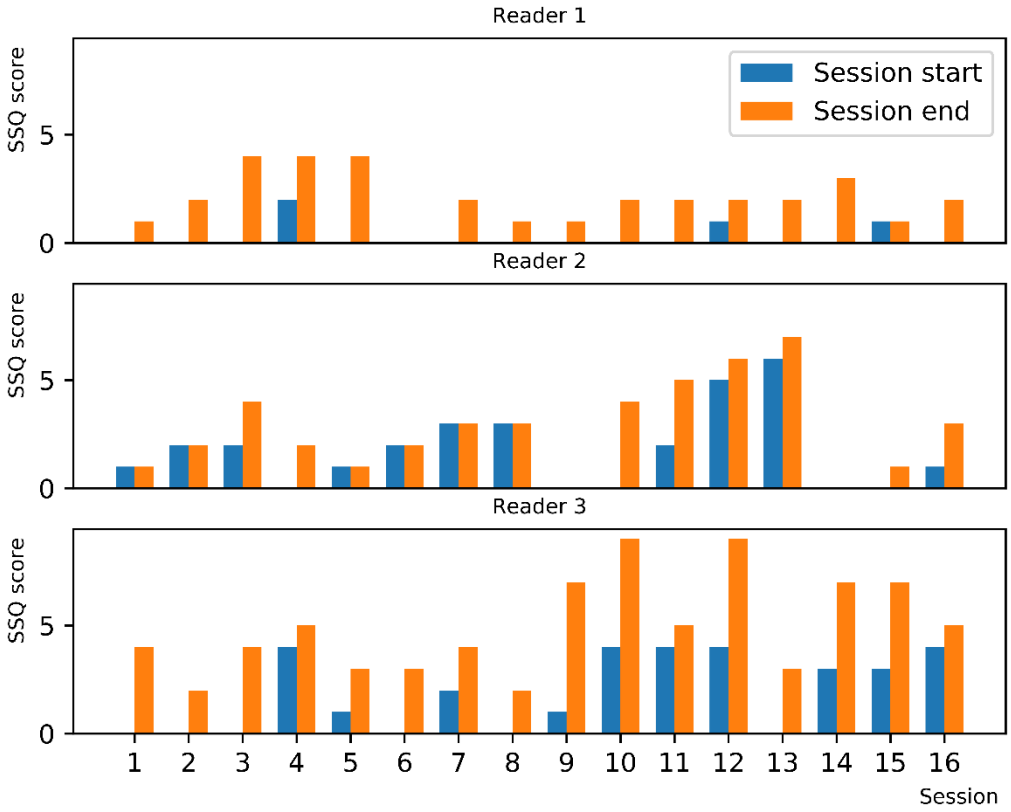


Figure 3-5. Reader discomfort per radiographer for annotating VFA images. Scores are given for the start and end of each reading session of six images. SSQ: Simulator Sickness Questionnaire.

3.6. References

1. Kanis JA on behalf of the World Health Organization Scientific Group (2007) Assessment of osteoporosis at the primary health-care level. Technical Report. World Health Organization Collaborating Centre for Metabolic Bone Diseases, University of Sheffield, UK. 2007. [Accessed 19-04-2021]
2. O'neill TW, Felsenberg D, Varlow J, Cooper C, Kanis JA, Silman AJ, European Vertebral Osteoporosis Study Group (1996) The prevalence of vertebral deformity in European men and women: the European Vertebral Osteoporosis Study. J Bone Miner Res 11(7):1010-1018. https://doi.org/10.1002/jbmr.5650110719
3. Felsenberg D, Silman AJ, Lunt M et al (2002) Incidence of vertebral fracture in europe: results from the European Prospective Osteoporosis Study (EPOS). J Bone Miner Res 17(4):716-724. https://doi.org/10.1359/jbmr.2002.17.4.716
4. Wade SW, Strader C, Fitzpatrick LA, Anthony MS, O'Malley CD (2014) Estimating prevalence of osteoporosis: examples from industrialized countries. Arch Osteoporos 9(1):182. https://doi.org/10.1007/s11657-014-0182-3
5. Gullberg B, Johnell O, Kanis JA (1997) World-wide projections for hip fracture. Osteoporos Int 7(5):407-413. https://doi.org/10.1007/pl00004148
6. Ross PD, Ettinger B, Davis JW, Melton L, Wasnich RD (1991) Evaluation of adverse health outcomes associated with vertebral fractures. Osteoporos Int 1(3):134-140. https://doi.org/10.1007/bf01625442

7. Melton LJ, Atkinson EJ, Cooper C, O'Fallon WM, Riggs BL (1999) Vertebral fractures predict subsequent fractures. Osteoporos Int 10(3):214-221. https://doi.org/10.1007/s001980050218
8. Black DM, Arden NK, Palermo L, Pearson J, Cummings SR (1999) Prevalent vertebral deformities predict hip fractures and new vertebral deformities but not wrist fractures. Study of Osteoporotic Fractures Research Group. J Bone Miner Res 14(5):821-828. https://doi.org/10.1359/jbmr.1999.14.5.821
9. Cauley JA, Hochberg MC, Lui LY, Palermo L, Ensrud KE, Hillier TA, Nevitt MC, Cummings SR (2007) Long-term risk of incident vertebral fractures. JAMA 298(23):2761-2767. https://doi.org/10.1001/jama.298.23.2761
10. Lindsay R, Silverman SL, Cooper C, Hanley DA, Barton I, Broy SB, Licata A, Benhamou L, Geusens P, Flowers K, Stracke H, Seeman E (2001) Risk of new vertebral fracture in the year following a fracture. JAMA 285(3):320-323. https://doi.org/10.1001/jama.285.3.320
11. Ismail AA, Cockerill W, Cooper C et al (2001) Prevalent vertebral deformity predicts incident hip though not distal forearm fracture: results from the European Prospective Osteoporosis Study. Osteoporos Int 12(2):85-90. https://doi.org/10.1007/s001980170138
12. Genant HK, Jergas M (2003) Assessment of prevalent and incident vertebral fractures in osteoporosis research. Osteoporos Int 14(Suppl 3):S43–55. https://doi.org/10.1007/s00198-002-1348-1
13. Roux C, Baron G, Audran M, Breuil V, Chapurlat R, Cortet B, Fardellone P, Trémollières F, Ravaud P (2011) Influence of vertebral fracture assessment by dual-energy X-ray absorptiometry on decision-making in osteoporosis: a structured vignette survey. Rheumatology (Oxford) 50:2264–9. https://doi.org/10.1093/rheumatology/ker225
14. Lewiecki EM, Laster AJ (2006) Clinical applications of vertebral fracture assessment by dual-energy x-ray absorptiometry. J Clin Endocrinol Metab *91*(11):4215-4222. https://doi.org/10.1210/jc.2006-1178
15. Malgo F, Hamdy NAT, Ticheler CHJM, Smit F, Kroon HM, Rabelink TJ, Dekkers OM, Appelman-Dijkstra NM (2017) Value and potential limitations of vertebral fracture assessment (VFA) compared to conventional spine radiography: experience from a fracture liaison service (FLS) and a meta-analysis. Osteoporos Int 28(10):2955-2965. https://doi.org/10.1007/s00198-017-4137-6
16. Shuhart CR, Yeap SS, Anderson PA, Jankowski LG, Lewiecki EM, Morse LR, Rosen HN, Weber DR, Zemel BS, Shepherd JA (2019) Executive summary of the 2019 ISCD position development conference on monitoring treatment, DXA cross-calibration and least significant change, spinal cord injury, peri-prosthetic and orthopedic bone health, transgender medicine, and pediatrics. J Clin Densitom 2*2*(4):453-471. https://doi.org/10.1016/j.jocd.2019.07.001
17. Genant HK, Wu CY, Van Kuijk C, Nevitt MC (1993) Vertebral fracture assessment using a semiquantitative technique. J Bone Miner Res *8*(9):1137-1148. https://doi.org/10.1002/jbmr.5650080915
18. Kennedy RS, Lane NE, Berbaum KS, Lilienthal MG (1993) Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. Int J Aviat Psychol *3*(3):203-220. https://doi.org/10.1207/s15327108ijap0303_3
19. Randolph JJ (2005) Free-Marginal Multirater Kappa (multirater K [free]): An Alternative to Fleiss' Fixed-Marginal Multirater Kappa. Paper presented at the Joensuu University Learning and Instruction Symposium 2005, Joensuu, Finland, October 14-15th, 2005
20. Koo TK, Li MY (2016) A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 15(2):155-163. https://doi.org/10.1016/j.jcm.2016.02.012

21. Davies KM, Recker RR, Heaney RP (1989) Normal vertebral dimensions and normal variation in serial measurements of vertebrae. J Bone Miner Res 4(3):341-349. https://doi.org/10.1002/jbmr.5650040308
22. Pearson D, Horton B, Green DJ, Hosking DJ, Goodby A, Steel SA (2006). Vertebral morphometry by DXA: a comparison of supine lateral and decubitus lateral densitometers. J Clin Densitom 9(3):295-301. https://doi.org/10.1016/j.jocd.2006.03.011
23. Bazzocchi A, Spinnato P, Fuzzi F, Diano D, Morselli-Labate AM, Sassi C, Salizzoni E, Battista G, Guglielmi G (2012) Vertebral fracture assessment by new dual-energy X-ray absorptiometry. Bone 50(4):836-841. https://doi.org/10.1016/j.bone.2012.01.018
24. Van Dort MJ, Romme EAPM, Smeenk FWJM, Geusens PPPM, Wouters EFM, van den Bergh JP (2018) Diagnosis of vertebral deformities on chest CT and DXA compared to routine lateral thoracic spine X-ray. Osteoporos Int, 29(6):1285-1293. https://doi.org/10.1007/s00198-018-4412-1
25. Fuerst T, Wu C, Genant HK, Von Ingersleben G, Chen Y, Johnston C, Econs MJ, Binkley N, Vokes TJ, Crans G, Mitlak BH (2009) Evaluation of vertebral fracture assessment by dual X-ray absorptiometry in a multicenter setting. Osteoporos Int 20(7):1199-1205. https://doi.org/10.1007/s00198-008-0806-9
26. Ferrar L, Jiang G, Schousboe JT, DeBold CR, Eastell R (2008) Algorithm-based qualitative and semiquantitative identification of prevalent vertebral fracture: agreement between different readers, imaging modalities, and diagnostic approaches. J Bone Miner Res 23(3):417-424. https://doi.org/10.1359/jbmr.071032
27. Damiano J, Kolta S, Porcher R, Tournoux C, Dougados M, Roux C (2006) Diagnosis of vertebral fractures by vertebral fracture assessment. J Clin Densitom 9(1):66-71. https://doi.org/10.1016/j.jocd.2005.11.002
28. Lee JH, Lee YK, Oh SH, Ahn J, Lee YE, Pyo JH, Choi YY, Kim D, Bae SC, Sung YK, Kim DY (2016) A systematic review of diagnostic accuracy of vertebral fracture assessment (VFA) in postmenopausal women and elderly men. Osteoporos Int, 27(5):1691-1699. https://doi.org/10.1007/s00198-015-3436-z
29. Cooper C, Atkinson EJ, O'Fallen WM, Melton LJ (1992) Incidence of clinically diagnosed vertebral fractures: a population based study in Rochester, Minnesota, 1985-1989. J Bone Miner Res 7:221-7. https://doi.org/10.1002/jbmr.5650070214
30. Delmas PD, van de Langerijt L, Watts NB, Eastell R, Genant HK, Grauer A, Cahall DL (2005) Underdiagnosis of vertebral fractures is a worldwide problem: the IMPACT study. J Bone Miner Res 20(4):557-63. https://doi.org/10.1359/jbmr.041214
31. Yang J, Cosman F, Stone PW, Li M, Nieves JW (2020) Vertebral fracture assessment (VFA) for osteoporosis screening in US postmenopausal women: is it cost-effective? Osteoporos Int 31(12): 2321-2335. https://doi.org/10.1007/s00198-020-05588-6

4. Initial validation of an Artificial Intelligence tool for automated Vertebral Fracture Assessment

In collaboration with P. Dibbets-Schneider, S.R. Romeijn, D.D.D. Rietbergen, L.M. Pereira Arias-Bouda, W. Grootjans

4.1. Abstract

Vertebral Fracture Assessment (VFA) is a standardized method for detecting and quantifying vertebral fractures. VFA is currently indicated for all patients who are at high risk of developing osteoporotic fractures. However, routine use of VFA has been proposed for detecting vertebral fractures in all patients referred for DXA imaging, though manual annotation of these VFA images is currently too labor-intensive and requires significant time investment of clinicians. Automation with AI-based software could assist readers to detect vertebral fractures on VFA images and may help to overcome these problems. In this pilot study, we evaluated the performance of a tool for the purpose of automating VFA in clinical routine.

Fifty-seven patients who underwent VFA imaging and spinal radiography of the thoracolumbar spine were retrospectively included. Genant classifications were automatically generated using the VFA images, and kappa scores for agreement with visual semiquantitative evaluation of spinal radiographs were calculated. A subset of twelve VFA images with good, moderate and poor automatic annotations was selected and manually corrected by three experienced radiographers. Correction time was recorded and compared to reader time without automated VFA, as well as inter-observer agreement.

Without manual correction, automated identification of patients with one or more vertebral fractures had a kappa score of 0.34 for agreement with visual fracture classification on spinal radiography for grade 1-3 and 0.43 for grade 2-3 fractures. The average difference in VFA annotation time with and without the automated annotation tool was 105 seconds for poor automatic annotations, 169 seconds for moderate annotations, and 231 seconds for good annotations.

Automated vertebral fracture assessment seems a viable approach to improve the detection of vertebral fractures. Although further improvements to the algorithm are needed, it shows the potential to significantly decrease annotation time.

4.2. Introduction

Osteoporosis causes around 9 million fractures worldwide each year. The average lifetime risk of sustaining an osteoporotic fracture of the wrist, hip or vertebra is about 30 to 40% in developed countries. Vertebral fractures are the most common osteoporotic fractures and have a major impact on patients' quality of life, through back-pain, reduced physical capability, increased need for daily care and reduced independence, perceived poor general health, and emotional status [1]. Incident vertebral fractures are associated with a marked deterioration of quality of life, and this deterioration is more pronounced in patients with a higher number of fractures [2, 3]. A vertebral fracture itself is a strong and independent predictor of subsequent fractures, not only of the spine but also the hip [4]. Adjusted for age and bone mineral density, a vertebral fracture is associated with a 4- to 5-fold increased risk of developing a subsequent vertebral fracture [5-7]. Timely diagnosis of vertebral fractures and effective treatment is thus of utmost importance to prevent further deterioration of health.

However, many vertebral fractures are not diagnosed, and these patients are therefore not adequately treated. About two thirds to three quarters of vertebral fractures are not clinically recognized [8, 9]. Many vertebral fractures are asymptomatic and may only be detected through spinal imaging. In

addition, even when presenting with symptoms, clinical signs and symptoms of vertebral fractures are not specific, and can easily be confused with other causes of back pain. Because of this, only an estimated 40% of elderly women and 20% of elderly men with vertebral fractures receive adequate anti-osteoporosis treatment [10-12].

Vertebral Assessment (VFA) using densitometry equipment is a convenient method to detect vertebral fractures since it can be performed simultaneously with measurement of bone mineral density and with a very low radiation dose to the patient. The International Society for Clinical Densitometry indicates VFA for all patients with bone density T-scores below -1.0 and one additional risk factor such as historical height loss, self-reported prior vertebral fracture, glucocorticoid use, or age ≥70 years for women or ≥80 for men [13]. For these indications, VFA and subsequent pharmaceutical treatment has been shown to be cost-effective [14]. Furthermore, calls have gone up to also routinely perform VFA in all patients visiting a Fracture Liaison Service (FLS), as the presence of vertebral fractures can provide more information for selecting the type and duration of treatment, but also to provide a baseline assessment to distinguish later incident vertebral fractures from prevalent fractures, important for treatment monitoring [15].

However, the application of VFA in clinical practice currently remains limited. The low image quality of VFA in comparison with conventional radiography complicates the objective classification of fractures, hence its relatively high inter-observer variability, even for quantitative assessment methods. In addition, quantitative assessment through vertebral morphometry requires a significant time investment from trained readers [16]. Automatic assessment of vertebral fractures may help to overcome both these limitations.

We are currently developing an AI-based software tool to assist readers to detect and classify vertebral fractures on VFA images. In this pilot study, we aim to apply this tool to assist readers for the first time to test our operating procedures and explore its viability. In addition, we evaluate the current performance of this software tool, and its potential impact on the clinical workflow. As the tool is still a prototype and under active development, we use these initial performance results as a proof of concept and as guidance for further development.

4.3. Methods

4.3.1. Data

VFA images included in a previous study performed by our group were used for this study. Inclusion criteria are described in detail in the corresponding report [16]. In short, 96 VFA images representing a wide range of age groups were selected, including both patients without vertebral fractures and patients with vertebral fractures of different types and severities on different vertebral levels. Of those, 57 patients underwent conventional lateral radiography of the thoracolumbar spine as well as VFA, and only these patients were included in the current study.

4.3.2. Image acquisition

Postero-anterior and lateral VFA images of the thoracolumbar spine (T4 - L4) were made by dedicated technicians using Hologic Horizon A DXA equipment (Hologic, Bedford, MA, USA. Software version 13.6). Patients were positioned in supine position with a cushion supporting the knees, and the c-arm was rotated for lateral imaging.

Lateral radiographs of the thoracic and lumbar spine were performed by a radiology technician using a standardized protocol, with the detector centralized on T7 for the thoracic spine and on L3 for the lumbar spine.

4.3.3. Assessment of vertebral fractures on radiographs

Radiographs were independently reviewed by a trained advanced practitioner and a radiologist. The semiquantitative Genant Classification scheme was used to assign grades to each vertebra from T4 to L4 [17]. Non-fractured vertebrae defined by a height loss less than 20% were assigned grade 0. Fractured vertebrae were assigned a grade between 1 and 3 based on the severity of the fracture: grade 1 for height loss between 20% and 25% (mild), grade 2 for a height loss between 25% and 40% (moderate), and grade 3 for a height loss of more than 40% (severe). Vertebrae that were considered not evaluable due to insufficient image quality, anomalies or other deformities were classified as not evaluable. In case of disagreements, the radiologist gave the final classification independently.

4.3.4. Manual assessment of vertebral fractures on VFA

Assessment of vertebral fractures on VFA images was done semi-automatically with 6-point morphometry using Hologic Physician Viewer software (Version 7.3, Hologic, Bedford, MA, USA). On each vertebra, three trained radiographers placed markers on the four corners of the two-dimensional projection of the vertebral body, and two markers in the middle of the upper and lower end plates respectively. The distance between upper and lower landmarks are used to calculate vertebral height ratios. Based on these morphological vertebral height measurements, vertebrae are classified according to the Genant classification scheme into normal, mild, moderate, or severe.

4.3.5. Automated assessment of vertebral fractures on VFA

Fully automatic detection and classification of vertebral fractures was done using the current alpha version of GECKO (Version 0.03.6, ImageBiopsy Lab, Vienna, Austria). This software package was developed by training a deep neural network on a dataset consisting of 2500 clinical VFA images from our center, with manual annotations made by experienced radiographers used as ground truth. GECKO autonomously locates the visible vertebrae in a VFA image and performs 6-point morphometry, similarly to manual assessment.

4.3.6. Standalone performance

The stand-alone diagnostic performance of automatic fracture assessment was evaluated per patient and per vertebra by comparing classifications made by GECKO to the visual classifications on spinal radiographs, which were considered the gold standard. Contingency tables were constructed, and accuracy, sensitivity, specificity and Cohen's Kappa were calculated as outcome measures of stand-alone performance. Because the clinical significance of grade 1 fractures is debated, analyses were done both including and excluding grade 1 vertebrae.

4.3.7. Combined human-machine performance

The GECKO software tool to automatically perform vertebral morphometry is not intended to replace human readers, but instead improve clinical workflow by decreasing the time needed to perform VFA. From all VFA images annotated by GECKO, a subset of 12 VFA images was selected to assess the potential clinical benefit of adding GECKO to the clinical workflow. Images with varying levels of accuracy of the initial automatic annotation were selected, four with good performance (12 correctly annotated vertebrae), four with moderate performance (9 correctly annotated vertebrae) and four with poor performance (6 correctly annotated vertebrae).

Automatic annotations were converted to DICOM format compatible with Hologic Physician Viewer software (Version 7.3, Hologic, Bedford, MA, USA). Three radiographers independently checked the automatic annotations and adjusted them when necessary. Since the same radiographers manually annotated these VFA images during the previous reader study, at least four months of time were between the first and second study to minimize memory effects. All images were presented to each radiographer in the same order, and were sequentially annotated in two sessions of six images. Between annotation sessions, a break of at least 15 minutes was planned. The time needed to annotate each image was recorded, and at the start and end of each session, radiographers filled in a modified Simulator Sickness Questionnaire (SSQ) to assess reader fatigue [18]. The readers were asked to score the presence of 7 common symptoms regarding fatigue and eyestrain on a 5-point scale (see appendix C); the overall SSQ score was given as the sum of these scores. Per-patient classification agreement with radiography was expressed using the Cohen's kappa score. Interobserver agreement between the three readers was assessed with the Randolph's kappa score.

4.4. Results

4.4.1. Standalone performance

Of the included patients, 40 (70%) were female and 17 were male, and the median age was 66 years (range 29 - 83). A total of 645 vertebrae were annotated in 57 VFA images, and 96 (13.0%) of the vertebrae between T4 and L4 were not annotated. An example annotation is shown in Figure 4-1.

On a per-patient level, GECKO correctly identified 27 of the patients with one or more vertebral fractures of any grade. Contingency tables are shown in Table 4-1-A. Only one fracture was missed, but 18 patients were incorrectly classified as fractured. Correspondingly, sensitivity was 96% at a specificity of 38%, with an accuracy of 67% and a Cohen's kappa of 0.34. Including only grade 2 and 3 fractures in this analysis resulted in an accuracy of 74%, sensitivity of 75% at a specificity of 73%, and a Cohen's kappa of 0.43 (see Table 4-1-B).

When considering grades 1 - 3 as fractured, 536 vertebrae were correctly identified as not fractured and 22 fractured vertebrae were correctly identified. Fifty-six vertebrae were incorrectly graded as fractured, and 31 fractures were missed (Table 4-2-A). The corresponding accuracy was 87%, with a sensitivity of 42%, specificity 91%, and a Cohen's kappa score of 0.26.

For grade 2 - 3 fractures only, 605 vertebrae were correctly identified as not fractured. Twelve fractured vertebrae were not identified as such, 22 normal vertebrae were incorrectly classified as fractured, and only 6 were correctly identified as fractured (Table 4-2-B). This resulted in 95% accuracy, and a sensitivity of 33%, specificity 96% and kappa 0.23.

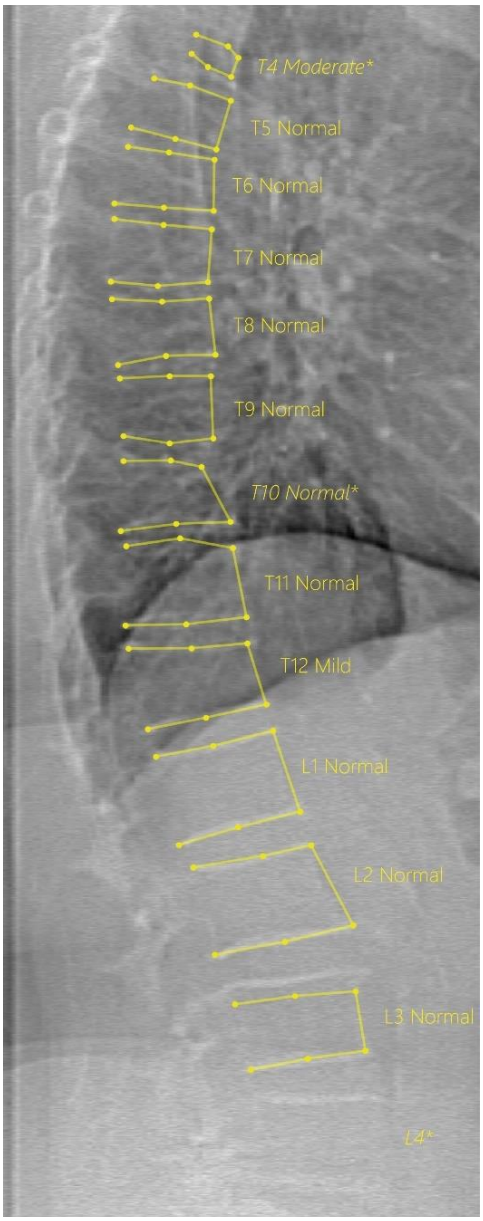


Figure 4-1. Example VFA with automatic annotation. L4 is not annotated. T4 is false positively classified as fractured. The mild fracture in T10 is missed. L3 shows incorrect localization of the upper endplate, but is correctly classified as normal.

Table 4-1. Contingency tables for automated identification of patients with one or more vertebral fractures on VFA images using GECKO; visual assessment on spinal radiographs as the golden standard. (A) Including fracture grades 1 - 3; (B) Including grade 2 and 3 fractures only.

Per patient					
A. Grade 1 - 3			**B.** Grade 2 - 3		
	GECKO not fractured	GECKO fractured		GECKO not fractured	GECKO fractured
DX not fractured	11	18	DX not fractured	30	11
DX fractured	1	27	DX fractured	4	12

DX: Spinal radiograph

Table 4-2. Contingency tables for automated identification of fractured vertebrae on VFA images using GECKO; visual assessment on spinal radiographs as the golden standard. (A) Including fracture grades 1 - 3; (B) Including grade 2 and 3 fractures only.

Per vertebra					
A. Grade 1 - 3			**B.** Grade 2 - 3		
	GECKO not fractured	GECKO fractured		GECKO not fractured	GECKO fractured
DX not fractured	536	56	DX not fractured	605	22
DX fractured	31	22	DX fractured	12	6

DX: Spinal radiograph

4.4.2. AI-assisted reader performance

4.4.2.1. Annotation time

The annotation time and GECKO correction time for each reader are shown in Figure 4-2. Without GECKO, annotation time ranged between 153 and 480 seconds, with an average annotation time of 274 seconds. For the images annotated by GECKO, the average time needed to correct an annotation was 155 (range 92 - 255) seconds for poor annotations, 120 (range 50 - 167) seconds for moderate annotations, and 66 (range 18 - 130) seconds for good annotations.

The average difference in annotation time with and without GECKO was 103 seconds for poor annotations, 156 seconds for moderate annotations, and 221 seconds for good annotations (see Table 4-3 and Figure 4-3).

One reader generally required shorter correction time than the others, with an average correction time of 73 seconds (95% CI: 51 - 94) compared to 139 seconds (95% CI: 103 - 174) and 130 seconds (95% CI: 94 - 165).

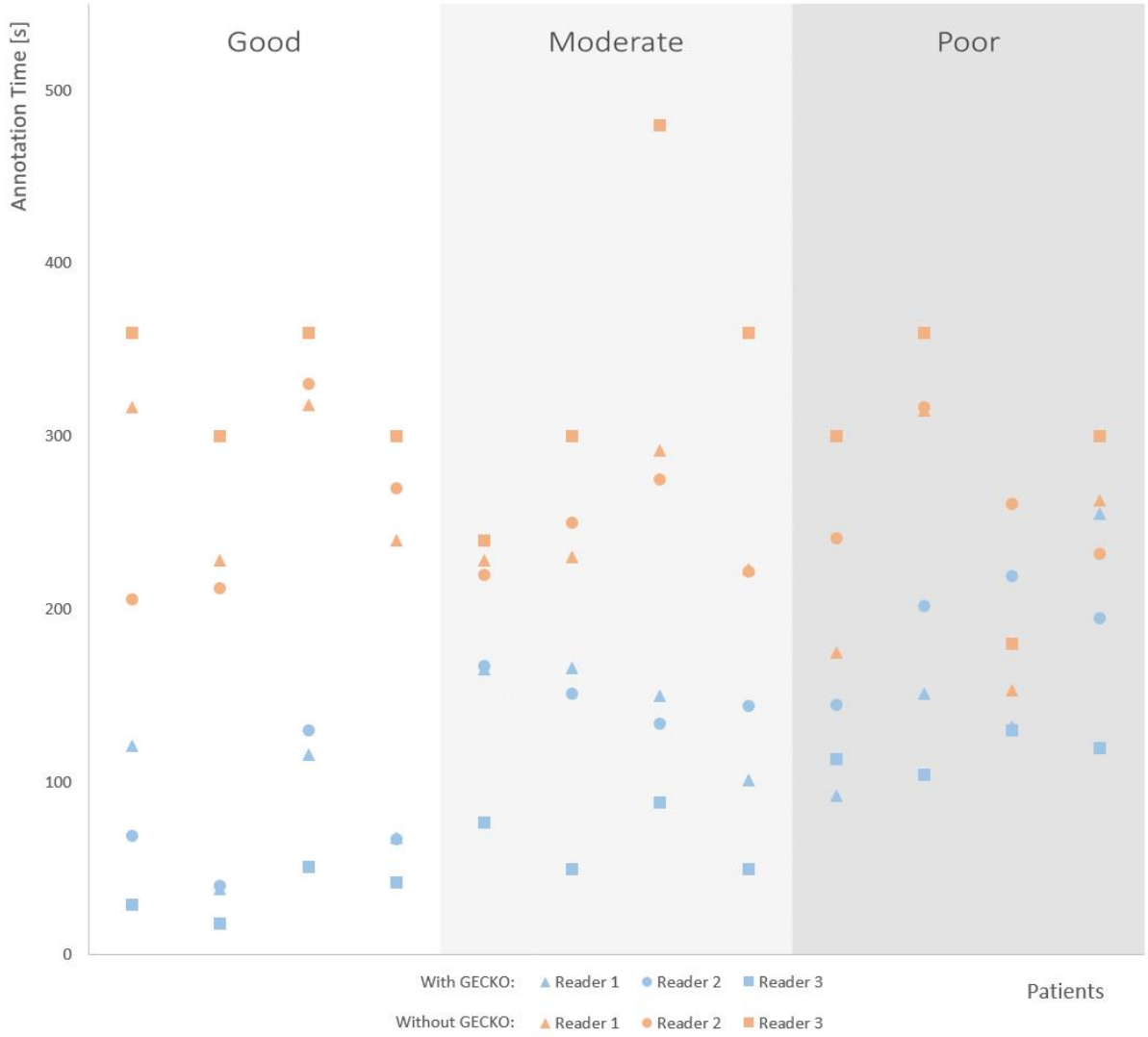


Figure 4-2. Annotation times with and without GECKO for good, moderate, and poor automatic annotations.

4.4.2.2. Reader Fatigue

SSQ scores measured at the start of a reading session ranged between 0 and 1, and scores measured at the end of a session ranged between 0 and 2. The average SSQ score was 0.33 at the start of a reading session and 0.83 at the end of a session (p=0.08).

4.4.2.3. Agreement with radiography

For the 12 selected patients, 3 patients had one or more grade 2 or 3 vertebral fractures diagnosed on radiography, and 3 additional patients had only grade 1 fractures. Only three patients with vertebral fractures were identified on VFA, resulting in an average kappa agreement score of 0.56 for grade 2-3 fractures and 0.46 for grade 1-3 fractures. We considered diagnosis on radiography the golden standard, and as such the sensitivity and specificity of the combined GECKO-radiographer VFA approach were 56% and 95% respectively.

4.4.2.4. Inter-observer agreement

Per-patient vertebral fracture identification was highest when including only grade 2 and 3 fractures, with a Randolph's kappa score of 0.73. When grade 1 vertebrae were also considered fractured, the kappa score was 0.33. Of the 156 vertebrae in the dataset, 5 (3%) were considered not evaluable by at least one of the radiographers. Randolph's kappa for agreement between radiographers for per-vertebra fracture severity classification was 0.89. Example images with automatic annotations and the corrections are shown in Appendix D.

Table 4-3. Mean annotation time [seconds] of VFA images with and without GECKO support for good, moderate and poor GECKO annotations.

	Without GECKO	With GECKO	Difference
Good	286.8	65.8	221.0
Moderate	276.7	120.3	156.4
Poor	258.1	154.8	103.3
Mean	273.8	113.6	

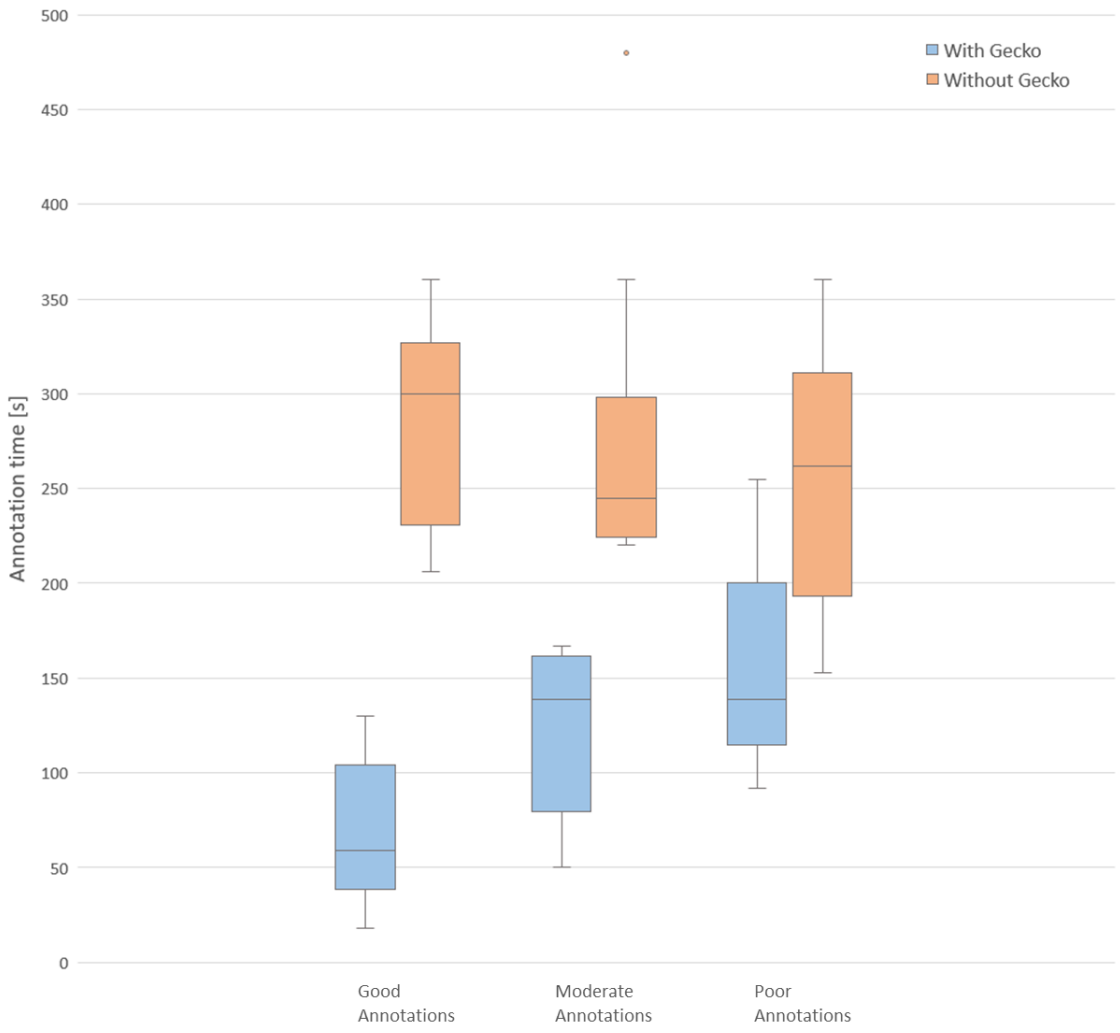


Figure 4-3. Annotation time boxplot with and without GECKO for good, moderate, and poor automatic annotations.

4.5. Discussion

4.5.1. Study findings

In this pilot study, we evaluated automated VFA fracture assessment with manual correction when necessary. In its current version, the automatic model did not reach good standalone performance, and had poor agreement with spinal radiography. However, when applied in conjunction with trained radiographers who corrected the automatic annotation when needed, annotation time was drastically reduced.

Standalone performance of the AI-based tool for identification of vertebral fractures was poor, with low agreement with radiography both for identification of patients with vertebral fractures of any grade and for grade 2-3 fractures only. In its current state, the model had a very high sensitivity, since only one patient with a vertebral fracture was missed. However, it seems that the model was not able to accurately identify patients without any vertebral fractures, resulting in a low specificity for fractures of any grade. Only including grade 2 and 3 fractures yielded a much higher specificity, but counterintuitively sensitivity decreased. A possible explanation for this is that the model did not annotate a relatively large number of vertebrae (13%), of which post-hoc analysis showed that a disproportionately large part was considered fractured on radiography. These fractures were then missed, and patients were incorrectly classified as not fractured, decreasing sensitivity.

Analyzing classifications per vertebra, classification agreement with radiography was very low (kappa 0.26 for grade 1-3 and 0.23 for grade 2-3 fractures), further showing that the current model's standalone performance is insufficient. Specificity however was high, mainly due to the intrinsic high proportion of normal vertebrae, which were generally correctly identified as such, as also indicated by the high accuracy scores. Sensitivity on the other hand was low, since many fractured vertebrae were missed and incorrectly classified as normal. Both sensitivity and specificity in the per-vertebra analyses seem to be distorted by the high proportion of normal vertebrae, and should be carefully interpreted. In absolute terms, in standalone analyses, GECKO missed 31 of the 53 fractures, and incorrectly classified another 56 normal vertebrae as fractured. It is clear that GECKO requires significant improvements. Further analysis of the annotations of this pilot study are to be used as ground for further training of the algorithm and improvement of post-processing. For example, it appears that annotations of fractured vertebrae are of lower quality than those of normal vertebrae, indicating that training with a focus on fractured vertebrae may be helpful.

Our annotation model is not intended to function completely autonomously, and automated annotations will have to be reviewed and corrected by human readers. Even though standalone performance of the automated VFA annotation model is not yet optimal, in this study we set out to evaluate the potential change in annotation time that the model may achieve when used in conjunction with human readers. As such, we selected a small subset of VFA images with different levels of quality of automated annotations. On this subset, this setup already showed an average time saving of 160 seconds per VFA image compared to the conventional semi-automatic annotation approach. As expected, time savings were much larger for automatic annotations where more vertebrae were annotated correctly, with an average time saving of 221 seconds for automatic annotations considered good, 156 seconds for moderate annotations, and 103 seconds for poor annotations. Surprisingly, even for poor annotations the average correction time was lower than the average annotation time without the assistance of automated annotations. These results show that supporting readers with automated VFA annotations is viable, even when annotations are not perfect.

These results also indicate that the largest improvements can be achieved with better automated annotations, further supporting the notion that GECKO should be improved. Although it is unlikely that GECKO will be able to perfectly annotate all VFAs in the near future, we believe its performance can still be significantly improved.

Agreement with radiography of GECKO with manual correction found in this study is similar to agreement scores of VFA without GECKO, with kappa scores ranging between 0.51 and 0.60 [16]. Inter-observer agreement across the three readers for identification of patients with one or more fractures was 0.73 for grade 2-3 fractures. Without GECKO, this was 0.65, slightly lower than with GECKO. However, direct comparison and statistical analysis is not possible with this small sample size. For grade 1-3 fractures, the kappa score was 0.33, showing that there was poor agreement between readers, mostly related to classification of grade 1 fractures. Grade 1 fractures are the most difficult to detect and subject to the highest variability [19], and the clinical relevance of mild fractures is still subject to debate. Nevertheless, in this pilot study GECKO has not been able to improve inter-operator agreement for grade 1 vertebral fractures.

4.5.2. Reader satisfaction

All participating radiographers reported that they found the workflow used in this study enjoyable. Although it was clear that automatic annotations were often inadequate, checking and correcting the initial annotations was considered an easier and less strenuous task than manually annotating an image completely. This was reflected in the SSQ scores, which remained low after reading sessions. All readers report that they would prefer working with GECKO over the current workflow.

Surprisingly, one reader required a significantly lower correction time than the other readers, while there was no evident difference for manual annotation. It might be that some readers more easily trust the AI annotations, while others are more skeptical and take longer to check the annotations.

4.5.3. Potential impact

In literature, studies on AI often report only standalone performance, often competing with other algorithms for achieving the highest possible standalone performance. Although in many cases standalone performance is important, it is often unclear what the added value to clinical practice is. For example, an algorithm can have near-perfect accuracy and still be of no added value if it fails to deliver what users need. On the other hand, algorithms need not always be perfect to be able to add value to clinical practice.

We believe that investigating the value and working closely with the intended users throughout the development process is essential. In this pilot study we aimed to investigate the potential added value of GECKO as part of its development process. Although the results show that further improvements are needed, it shows the potential to significantly decrease annotation time and decrease reader fatigue. This could make it possible to perform more VFAs in a certain amount of time, and potentially increasing the number of detected vertebral fractures, allowing more patients to be adequately treated.

4.5.4. Limitations

This study has some important limitations. Firstly, its small sample size makes it hard to generalize the outcomes. As such, the results should be interpreted with caution. However, this pilot study was intended as an initial investigation and its results are to be confirmed in a larger study in a later stage.

In addition, readers were asked to correct the GECKO annotations using Physician's Viewer software, which is currently also used in clinical practice. However, in the future GECKO annotations will be

corrected in a user interface specifically designed for this purpose, and a different interface could affect annotation times. In addition, in this study annotation time measurements were started when readers started correcting the automatic annotations. That means that in the clinical workflow, automatic annotations need to be readily available to the readers to achieve the potential time savings. Any delays or additional actions that readers need to perform can decrease this time saving.

Participating radiographers were very experienced in VFA annotation. Readers with less experience may require longer annotation times. However, readers had no experience with GECKO and no additional training was provided for reviewing automatic annotations. With more training or experience, annotation times might further improve.

The distinction between poor, moderate, and good annotations made for this investigation is relatively arbitrary. It was based on the number of correct annotations, but there was no upper or lower limit on the number of incorrect annotations. Some poor annotations had six correctly annotated vertebrae and one or two incorrectly annotated, while others had six correct and six incorrect annotations, which could have affected correction time.

SSQ scores are subjective and few measurements could be acquired because only two reading sessions per reader were held. Therefore, statistical analysis was not possible. In later studies subjective findings could be supported by quantitative measurements of discomfort by measuring biometric data such as eye movement or muscle tension. In addition, other measures of reader effort besides annotation time could be added, such as the number of mouse clicks.

Finally, we used vertebral fracture classification on spinal radiography as the golden standard. Although this is the best information available, there is no true gold standard. As such, patients' true fracture status is unknown, complicating the interpretation of the performance metrics for identification of vertebral fractures on VFA [19].

4.5.5. Conclusion

Automated vertebral fracture assessment seems a viable approach to improve the detection of vertebral fractures. Supporting readers with automatic annotations for vertebral morphometry shows the potential to significantly decrease reader time and decrease reader fatigue, but further improvements to the algorithm are needed.

4.6. References

1. Ross PD, Ettinger B, Davis JW, Melton L, Wasnich RD (1991). Evaluation of adverse health outcomes associated with vertebral fractures. Osteoporos Int. 1(3):134-140. https://doi.org/10.1007/bf01625442
2. Oleksik A, Lips P, Dawson A, Minshall ME, Shen W, Cooper C, Kanis J (2000). Health-related quality of life in postmenopausal women with low BMD with or without prevalent vertebral fractures. J Bone Miner Res. 15:1384-1392.
3. Silverman SL, Minshall ME, Shen W, Harper KD, Xie S (2001). The relationship of health-related quality of life to prevalent and incident vertebral fractures in postmenopausal women with osteoporosis: results from the Multiple Outcomes of Raloxifene Evaluation Study. Arthritis Rheum. 44:2611-2619.
4. McCloskey EV, Vasireddy S, Threlkeld J, Eastaugh J, Parry A, Bonnet N, Beneton M, Kanis JA, Charlesworth D (2008). Vertebral fracture assessment (VFA) with a densitometer predicts future fractures in elderly women unselected for osteoporosis. J Bone Miner Res. 23:1561-1568.
5. Black DM, Arden NK, Palermo L, Pearson J, Cummings SR (1999). Prevalent vertebral deformities predict hip fractures and new vertebral deformities but not wrist fractures. Study of Osteoporotic Fractures Research Group. J Bone Miner Res. 14:821-828.
6. Cauley JA, Hochberg MC, Lui LY, Palermo L, Ensrud KE, Hillier TA, Nevitt MC, Cummings SR (2007). Long-term risk of incident vertebral fractures. JAMA. 298:2761-2767.
7. Lindsay R, Silverman SL, Cooper C, Hanley DA, Barton I, Broy SB, Licata A, Benhamou L, Geusens P, Flowers K, Stracke H, Seeman E (2001). Risk of new vertebral fracture in the year following a fracture. JAMA. 285(3): 320-323.
8. Cooper C, O'Neill T, Silman A (1993). The epidemiology of vertebral fractures. European Vertebral Osteoporosis Study Group. Bone 14 Suppl 1:S89-97.
9. Fink HA, Milavetz DL, Palermo L, Nevitt MC, Cauley JA, Genant HK, Black DM, Ensrud KE (2005) What proportion of incident radiographic vertebral deformities is clinically diagnosed and vice versa? J Bone Miner Res 20:1216-1222.
10. Majumdar SR, Kim N, Colman I, Chahal AM, Raymond G, Jen H, Siminoski KG, Hanley DA, Rowe BH (2005). Incidental vertebral fractures discovered with chest radiography in the emergency department: prevalence, recognition, and osteoporosis management in a cohort of elderly patients. Arch Intern Med. 165:905-909.
11. Kroth PJ, Murray MD, McDonald CJ (2004). Undertreatment of osteoporosis in women, based on detection of vertebral compression fractures on chest radiography. Am J Geriatr Pharmacother 2:112-118.
12. Panneman MJ, Lips P, Sen SS, Herings RM (2004). Undertreatment with anti-osteoporotic drugs after hospitalization for fracture. Osteoporos Int. 15:120-124.
13. Shuhart CR, Yeap SS, Anderson PA, Jankowski LG, Lewiecki EM, Morse LR, Rosen HN, Weber DR, Zemel BS, Shepherd JA (2019) Executive summary of the 2019 ISCD position development conference on monitoring treatment, DXA cross-calibration and least significant change, spinal cord injury, peri-prosthetic and orthopedic bone health, transgender medicine, and pediatrics. J Clin Densitom 22(4):453-471.
14. Schousboe, J. T., Ensrud, K. E., Nyman, J. A., Kane, R. L., & Melton III, L. J. (2006). Cost-effectiveness of vertebral fracture assessment to detect prevalent vertebral deformity and select postmenopausal women with a femoral neck T-score>− 2.5 for alendronate therapy: a modeling study. Journal of Clinical Densitometry, 9(2), 133-143.

15. Lems WF, Paccou J, Zhang J, Fuggle NR, Chandran M, Harvey NC, Cooper C, Javaid K, Ferrari S, Akesson KE (2021). Vertebral fracture: epidemiology, impact and use of DXA vertebral fracture assessment in fracture liaison services. Osteoporos Int. 32(3):399-411. https://doi.org/10.1007/s00198-020-05804-3
16. Mostert JM, Romeijn SR, Dibbets-Schneider P, Götz C, DiFranco MD, Dimai HP, Grootjans W (2021). Inter-operator agreement of vertebral fracture assessment with dual-energy x-ray absorptiometry equipment. This thesis, chapter 3.
17. Genant HK, Wu CY, Van Kuijk C, Nevitt MC (1993). Vertebral fracture assessment using a semiquantitative technique. J Bone Miner Res. *8*(9):1137-1148. https://doi.org/10.1002/jbmr.5650080915
18. Kennedy RS, Lane NE, Berbaum KS, Lilienthal MG (1993). Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. Int J Aviat Psychol. 3(3):203-220. https://doi.org/10.1207/s15327108ijap0303_3
19. Oei L, Koromani F, Breda SJ et al (2018). Osteoporotic vertebral fracture prevalence varies widely between qualitative and quantitative radiological assessment methods: the Rotterdam Study. J Bone Miner Res. 33(4):560-568.

5. General discussion

User-centered co-development of an Artificial Intelligence application for automated vertebral fracture assessment

5.1. Introduction

Artificial Intelligence (AI) has been one of the main research topics in radiology in recent years, and the number of AI products on the market has rapidly grown as well. To date, over 100 AI products have received CE-marking for market access in Europe [1], more than 120 have been cleared by the FDA for clinical use in the United States [2], and more than 10-fold growth in the market in the next decade is predicted [3]. Many radiologists also have a positive attitude towards AI and seem willing to adopt AI products [4-6]. However, clinical implementation of AI in radiology has been slow and currently remains limited [7].

A number of factors have been identified as barriers to widespread adaptation of AI in radiology, including inconsistent technical performance of AI products, unstructured implementation processes, uncertain added value for clinical practice, large variance in acceptance and trust of its users, legal and ethical constraints, and poor workflow integration [4, 8, 9]. Although AI has the potential to transform the field of radiology, overcoming these barriers is paramount to its effective adaptation in clinical practice. For the majority of AI products available on the market, there is no peer-reviewed evidence of its efficacy. Even when scientific literature is available, it is often limited to evaluating technical performance, while added value to clinical practice is rarely investigated [1]. Some argue that this is because the field of AI in radiology is still in its early stages, and it is a matter of time before more studies can investigate the added value to clinical practice. However, we believe investigations like that can and should also be performed in the early development stages of AI products. To do that, it is essential for academic healthcare institutions and industry to work together in the development of AI products. In this opinion article we describe our process for user-centered co-development of an AI application for automated vertebral fracture assessment, and reflect on the early-stage investigations into its potential clinical value.

5.2. User-centered co-development

The concept of co-development encompasses the idea that suppliers of medical technology work in very close partnership with healthcare providers, clinicians and clinical researchers to develop new med-tech products. This allows the industry to tackle problems that could not be addressed by either medical technology suppliers or healthcare providers alone.

The general idea behind this co-development strategy is to utilize the strengths of multiple stakeholders to supplement each other in terms of access to knowledge, data, and other resources that are required to develop successful AI products. And, perhaps most importantly, co-development aims to ensure that clinical need remains central to the development process and that an existing clinical problem is tackled. At the 2019 meeting of the International Society for Strategic Studies in Radiology, the most important ingredients for success were discussed, as well as access to these ingredients for the largest stakeholders for developing AI tools (see Figure 5-1) [10]. Based on each stakeholder's strengths and weaknesses, partnerships between academic healthcare institutions and industry seem the most promising, since together they have access to all resources needed for successful development of AI applications.

Partnerships between healthcare institutions and medical technology companies are very common. In fact, many would argue that partnerships in some form are essential to the development of new products. However, the extent of partnerships may differ. For example, a medical device manufacturer may only seek out a collaboration with a hospital for a validation study of a product that was developed independently. Development of AI products is generally bound to another type of collaboration to acquire training data, since companies are usually not able to generate clinical data themselves. Co-development is not a clearly defined term, but as the name already suggests, it involves intensive collaboration between partners throughout all product development phases. In that sense, co-development goes beyond conventional partnerships.

In addition, the term 'user-centered' indicates that usability, workflow, and user experience are emphasized throughout the development process. User-centered development attempts to optimize a product around how end-users prefer to use it. As such, intended users - whether radiologists, radiographers, or technicians - are actively involved in the development process.

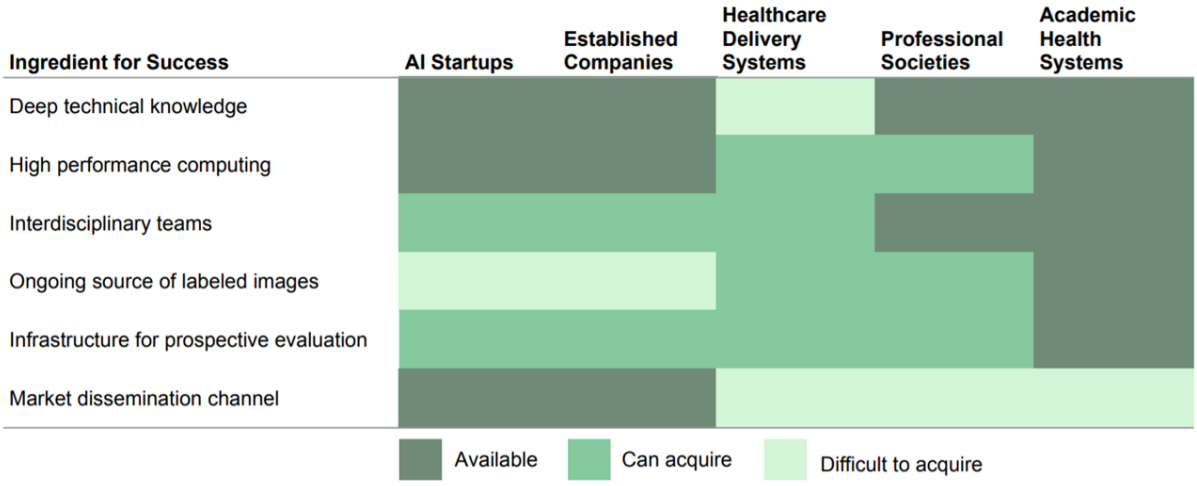


Figure 5-1. Ingredients for successful development of AI products and major stakeholders' access to the ingredients for success. Adapted from Recht et al. [10].

5.3. The AI design cycle

There are many formal methodologies for the development of medical devices described in literature [11]. Each method defines a number of distinct development stages, from concept to discontinuance. Although each method defines slightly different stages, in essence each method contains a process for translating design inputs (e.g., user needs, regulatory requirements) to design outputs (the product). Design outputs then enter the verification stage or validation stage, where verification internally checks if the design output fulfills its specifications, and validation externally checks if the product meets the user needs. We adopted a basic iterative design process in which verification and validation are centrally integrated. As shown in Figure 5-2,the design cycle consists of five distinct processes which are to be followed in order and can be repeated as many times as necessary. Outputs of verification and validation refer back to the design input and are used to adjust the design specifications as needed.

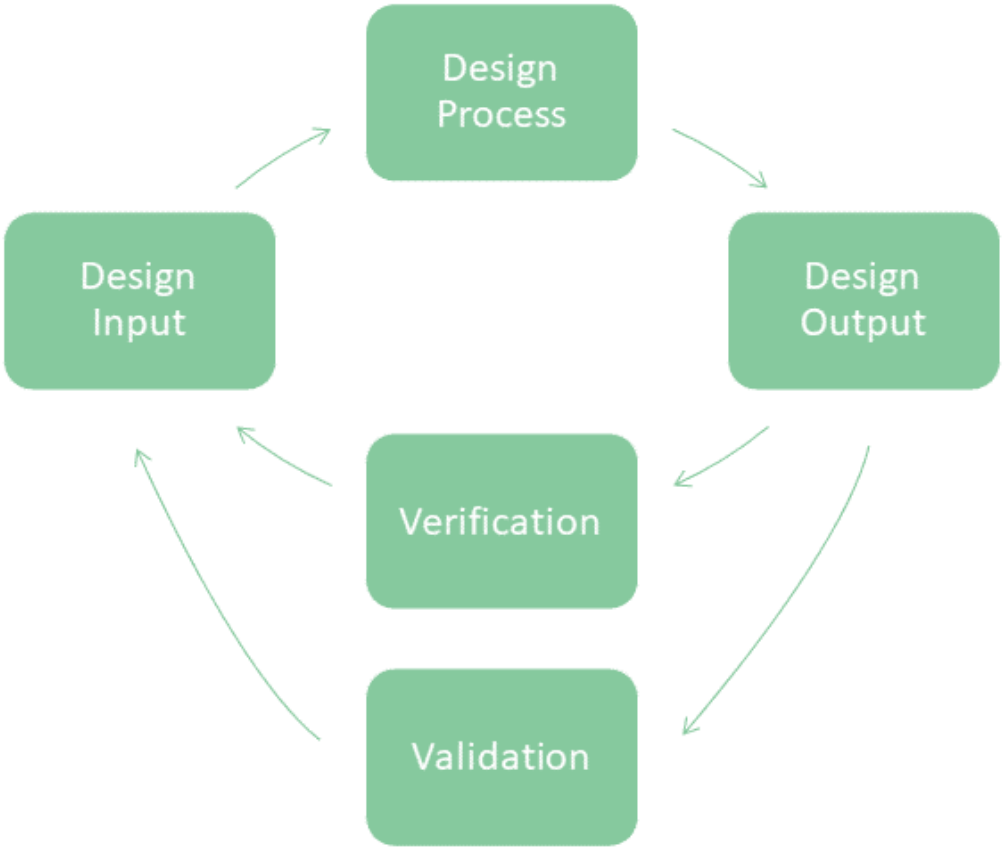


Figure 5-2. Schematic description of the iterative design process for Artificial Intelligence applications.

5.4. Validation types

Validation of AI applications is an essential part of its development process and of great importance for market access. There are multiple types of validation, which can be classified in a hierarchical model of efficacy. In a recent paper, Van Leeuwen et al. proposed a hierarchical model specifically for AI software for diagnostic imaging [1]. It consists of six levels from technical efficacy (level 1t for technical feasibility and 1c for clinical feasibility) up to societal impact (level 6). A detailed description of all levels is given in Table 5-1. Validation of new AI applications generally starts at a low level, and moves up to higher levels of validation over time. There is no clear cut-off for defining what level of validation is necessary for an application to be accepted in clinical practice. Some applications do not require more than a description of its stand-alone diagnostic accuracy, while others require higher validation levels. However, within value-based healthcare systems, the primary goal of innovations is to add value to the healthcare system. The hierarchical model described here provides a framework relating the level of validation with its value to the healthcare system.

Table 5-1. Hierarchical model of efficacy to assess the contribution of AI software to the diagnostic imaging process. Adapted from van Leeuwen et al. (2021) [1].

Level	Explanation	Typical measures
Level 1t	Technical efficacy Article demonstrates the technical feasibility of the software	Reproducibility, inter-software agreement, error rate
Level 1c	Potential clinical efficacy Article demonstrates the feasibility of the software to be clinically applied	Correlation to alternative methods, potential predictive value, biomarker studies
Level 2	Diagnostic accuracy efficacy Article demonstrates the stand-alone performance of the software	Standalone sensitivity, specificity, area under the ROC curve, or Dice score
Level 3	Diagnostic thinking efficacy Article demonstrates the added value to the diagnosis	Radiologist performance with/without AI, change in radiological judgement
Level 4	Therapeutic efficacy Article demonstrates the impact of the software on patient management decisions	Effect on treatment or follow-up examinations
Level 5	Patient outcome efficacy Article demonstrates the impact of the software on patient outcomes	Effect on quality of life, morbidity, or survival
Level 6	Societal efficacy Article demonstrates the impact of the software on society by performing an economic analysis	Effect on costs and quality-adjusted life years, incremental costs per quality-adjusted life year

Level 1t level 1, technical; *Level 1c* level 1, clinical

5.5. Automated Vertebral Fracture Assessment

5.5.1. The idea

Most innovations start out with an initial problem, for which someone wants a solution. Oftentimes - in our case as well - the idea comes from a clinician, who has to deal with a certain problem in clinical practice. In our case, radiographers do not enjoy performing vertebral fracture assessment. This strenuous task, in which landmarks are placed on vertebral bodies on a lateral spine image, requires a high degree of focus and detail for attention. This makes vertebral fracture assessment a repetitive, time- and labor-intensive task. During this annotation task, readers are supported by an algorithm, but this algorithm rarely works correctly. Thus, one of the radiographers decided that this can be improved, and sought out an industry co-developer to develop the idea into a functional product. Together, we employed a user-centered co-development strategy aiming to create an AI tool to automatically perform vertebral morphometry for vertebral fracture assessment.

5.5.2. The available tools

With an extensive literature search [12], we explored the current application of AI on fracture risk assessment using Dual Energy X-ray Absorptiometry (DXA) equipment, the same equipment used for VFA. DXA can be used to determine several parameters that can be used to quantify a patient's risk of developing a vertebral fracture. The presence of vertebral fractures is only one of those parameters, and notable others are bone mineral density, trabecular bone score, and hip geometry. AI can be applied to aid in the extraction of these parameters from DXA images, as well as multivariate fracture risk prediction to support clinical decision making. Several studies report on automating the detection of vertebral fractures on VFA images, showing good performance. However, few tools are available on the market, and the number of investigations into their clinical value is very limited.

5.5.3. Quantification of current methods

Current semi-automatic quantitative vertebral fracture assessment methods are far from perfect. Multiple studies indicate that its inter-observer variability is relatively high [13-15]. In addition, readers indicate that it is a labor-intensive task that requires relatively long reading time and has a detrimental effect on reader fatigue. We could not find any published literature reporting on VFA reading time or reader fatigue. Therefore, we performed a reader study to quantify the time and effort needed to manually perform vertebral morphometry [16]. This study showed that readers spent up to 540 seconds to annotate a VFA image, with an average annotation time of 260 seconds. Besides, readers report a significant increase in physical complaints related to fatigue and eyestrain after annotating VFA images. Finally, inter-observer agreement for identification and classification of vertebral fractures on VFA was only moderate. These results serve as a baseline measurement of the current practice, and clearly indicate that there is room for improvement.

5.5.4. Algorithm development

AI algorithms require large amounts of data, not only for development and algorithm training, but also for testing and validation. In the Leiden University Medical Center, over 3000 clinical VFAs are made each year. These VFAs can be used as a basis for the development of our AI algorithm, since they represent a large cohort of patients who underwent VFA for a wide range of indications. However, to be useful for training an AI algorithm, these images need to be accompanied by high-quality ground truth information. Vertebral morphometry of all vertebrae from T4 to L4 is part of the standard assessment protocol for VFA in our institution. Not only the vertebral height measurements and fracture classifications are saved, but also the annotations made for this quantitative assessment are stored in our Picture Archiving and Communication System (PACS). As such, these manual annotations can be used as the ground truth for training an AI algorithm to perform vertebral morphometry on VFA images.

We retrospectively selected all annotated VFA images made between January 1st 2019 and January 1st 2020 for training. Our co-development partner designed an AI algorithm to perform vertebral morphometry and trained it using the manually annotated VFA images. This algorithm was named GECKO. As stated above, development of an AI application is much more than simply creating a model and training it on the available data. It is an iterative process in which an initial prototype is developed, verified or validated, and then improved upon based on the outcome of the verification or validation. Here we describe some of our findings from the validation study we performed in this process.

5.5.5. Technical validation

As part of the technical validation, we performed an initial evaluation of the algorithm's standalone performance for vertebral fracture identification [17]. Although the product is not intended to function autonomously and will require human review and adjustment, its standalone performance is still an

important indicator of its potential clinical value. In a selected sample of 57 VFA's, we compared GECKO's vertebral classifications to the golden standard as determined by conventional radiography. As is usual for VFA images, some vertebrae are not sufficiently visible and cannot be evaluated. In our sample, GECKO did not annotate 13% of vertebrae between L4 and T4, only slightly more than trained radiographers (9.7%). However, the vertebrae that were evaluated were not often classified correctly. For grade 2 and 3 fractures, GECKO had a high specificity of 96%, but a very low sensitivity of 33%. Agreement with classification on radiographs was poor (kappa 0.23).

These findings confirmed our expectations that the algorithm does not function optimally yet, mainly missing vertebral fractures and incorrectly classifying them as normal. Visual assessment of the annotations made by GECKO did provide some insights in the cases where it often fails, for example lower lumbar vertebrae (mostly L4) are often not annotated, and upper thoracic vertebrae are often annotated incorrectly, most likely due to interfering lung patterns. We believe that with these insights we can further improve the algorithm performance. This will not only be done by increasing training volume, but also by improving internal logic with conventional image analysis to filter out obvious errors. Besides that, it is important to realize that perfect standalone performance is not required to add value to clinical practice. In this same sample, classification accuracy was 95% due to the algorithm's ability to correctly identify most normal vertebrae. Although annotating fractured vertebrae was suboptimal, annotation of normal vertebrae can still be useful to users, so they can focus on fractured vertebrae more easily.

5.5.6. Clinical validation

Although technical validation and clinical validation are often performed serially, this is not always necessary. On the contrary, performing clinical and technical validations in parallel allows us to evaluate how our application fits within the clinical workflow and affirm hypotheses regarding the potential value of an application, without having fully optimized its technical performance. Therefore, we performed an exploratory investigation into the potential clinical value of the GECKO.

Although we hypothesized that automated VFA annotation followed by manual correction would be faster than the current workflow, there was no available data to support that hypothesis. In a small-scale pilot study, we aimed to evaluate and quantify the potential time saving that could be reached when using GECKO. Automatic annotations were made and a careful selection was made including good, moderate, and poor annotations. Without any additional training, readers then checked the automatic annotations and manually adjusted them where necessary within a workflow simulating normal clinical practice. As we hypothesized, there was a clear decrease in annotation time when using GECKO compared to annotating without GECKO. While time savings were largest when automatic annotations were good (mean 221 seconds), there was still an average decrease in time of 103 seconds for poor annotations.

These results show that GECKO seems viable in clinical practice and fits within the workflow that users are already familiar with. Although we could not systematically assess reader satisfaction, all readers reported that they enjoyed working with GECKO and preferred it over the current annotation method.

5.5.7. Clinical Benefit Analysis

The primary driver behind the recent emergence of AI in radiology has been the need for more efficiency in clinical care. Radiological imaging data continues to grow at a disproportionate rate when compared with the number of available trained readers, and the decline in imaging reimbursements has forced health-care providers to compensate by increasing productivity [18]. With this, workload on radiologists has dramatically increased [19], potentially increasing the risk of errors. As such, there is a

large desire for tools to increase efficiency. Besides these quantitative outcomes, other qualitative parameters such as job satisfaction and enjoyability of work are also important to avoid radiologists and radiographers developing burn-out symptoms. AI can be used to perform many different types of tasks in medical imaging, usually involving detection, quantification, or monitoring of abnormalities. Automating labor-intensive, repetitive quantification tasks, such as segmentation or measurement of certain quantitative parameters, has been an important spearpoint in the development of AI in radiology.

Nevertheless, there is a severe lack of evidence supporting the notion that AI tools can improve clinical efficiency [1]. We believe that it is very important for all stakeholders to assess the clinical value of AI applications, and this can be done early in the development stage as well. For our application automating VFA, the potential benefit lies mainly with decreasing the annotation time needed to perform VFA. However, a single patient would not benefit from a faster annotation time, as annotations are made after a patient has left the examination room. The real benefit would be for the system as a whole, since decreasing the time needed to perform VFA means that more patients can undergo VFA without additional investment in more personnel or equipment.

In our Fracture Liaison Service (FLS), on average 20 patients undergo BMD measurement per day. Of these 20 patients, for only 10 patients VFA is performed. As a crude estimation, the current algorithm could achieve an average time saving of 160 seconds per VFA. This corresponds to a total of 26.7 minutes per day. With an average image acquisition time of 4 minutes and an annotation time of slightly more than 2 minutes, this means that 4 additional patients can undergo VFA each day, without any additional time investments. As more patients can undergo VFA, more subclinical vertebral fractures could be detected, allowing for more accurate treatment of those at risk of subsequent osteoporotic fractures.

The results of this small-sample pilot show that the time needed to annotate VFA images could drastically be reduced with the use of automated annotation software. As one might expect, the potential time savings are highly dependent on the quality of the initial automatic annotations. Although our automatic annotation tool is still under development and we believe its standalone performance can be improved, it is unlikely that GECKO will be able to perfectly annotate all VFAs in the near future. Therefore, verification and - if needed correction - by a human reader will remain necessary. Nevertheless, GECKO may have a significant impact on clinical practice. We believe it is possible to reach an average performance at the level considered ‘good’ in this study, annotating around 90% of vertebrae correctly. With this level of performance, even more patients can undergo VFA, and eventually all FLS patients can undergo VFA without any additional time required.

5.6. Next steps

Although the results of the initial investigations seem promising, we are far from finished. To make a clinically viable product, further improvement of the model performance is essential. Several proposed changes based on the outcomes from our investigations are currently being implemented and will soon be reevaluated.

Meanwhile, we have worked over the previous period to set up a large independent dataset that can be used for thoroughly validating the standalone performance of the updated algorithm. This dataset consists of almost 600 VFAs made using a different scanner model than the training data, has already been manually annotated, and is accompanied by classifications on conventional radiography that serves as the golden standard. Once the planned improvements to the algorithm have been implemented and satisfactorily evaluated, we will use this dataset for an independent validation of the

algorithm standalone performance. The goal is to eventually get the algorithm available on the market as a medical device software package, and set up prospective studies to further evaluate the added value to clinical practice.

In addition, we plan on re-using the validation strategy applied in this project for other automation applications as well. Our industry partner has developed multiple AI-based applications automating quantitative measurements on medical images, some of which are already available on the market in Europe and the United States. We are planning on expanding the partnership, setting up more prospective studies focusing on the value of these applications to clinicians.

5.7. Conclusion

Both industry and healthcare institutions can benefit from extensive collaboration. More importantly, these collaborations can lead to technical innovations that clinicians and patients can benefit from as well. Working together with healthcare professionals, AI algorithms can automate many tasks to decrease workload and improve patient care. Evaluating the added value to clinical practice remains essential. Automated vertebral fracture assessment seems feasible, and in the near future, readers can be supported by our AI-based application, potentially leading to lower annotation times and improving clinical workflow.

5.8. References

1. Van Leeuwen KG, Schalekamp S, Rutten MJ, Van Ginneken B, De Rooij M. (2021). Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. Eur Radiol. Apr 15:1-8.
2. ACR Data Science Institute. FDA-cleared AI algorithms. Available at: https://www.acrdsi.org/DSI-Services/FDA-Cleared-AI-Algorithms. Accessed 06-05-2021.
3. Data Bridge Market Research (2021). Global Artificial Intelligence in Medical Imaging Market – Industry Trends and Forecast to 2028. Available at: https://www.databridgemarketresearch.com/reports/global-artificial-intelligence-in-medical-imaging-market
4. Huisman M, Ranschaert E, Willemink MJ et al (2021). An international survey on AI in radiology in 1,041 radiologists and radiology residents part 1: fear of replacement, knowledge, and attitude. Eur Radiol. May 11:1-10.
5. Waymel Q, Badr S, Demondion X, Cotten A, Jacques T (2019). Impact of the rise of artificial intelligence in radiology: what do radiologists think?. Diagn Interv Imaging. 100(6):327-336.
6. Coppola F, Faggioni L, Grassi R et al (2021). Artificial intelligence: radiologists' expectations and opinions gleaned from a nationwide online survey. La Radiologia Medica. 126(1):63-71.
7. Allen B, Agarwal S, Coombs L, Dreyer K, Wald C (2021). 2020 ACR Data Science Institute Artificial Intelligence Survey. J Am Coll Radiol. S1546-1440(21)00293-3.
8. Strohm L, Hehakaya C, Ranschaert ER, Boon WP, Moors EH (2020). Implementation of artificial intelligence (AI) applications in radiology: hindering and facilitating factors. Eur Radiol. 30:5525-5532.
9. Huisman M, Ranschaert E, Willemink MJ (2021). An international survey on AI in radiology in 1041 radiologists and radiology residents part 2: expectations, hurdles to implementation, and education. Eur Radiol. May 11:1-10.
10. Recht MP, Dewey M, Dreyer K, Langlotz C, Niessen W, Prainsack B, Smith JJ (2020). Integrating artificial intelligence into the clinical practice of radiology: challenges and recommendations. Eur Radiol. 30(6):3576-3584.

11. Marešová P, Klímová B, Honegr J, Kuča K, Ibrahim WNH, Selamat A (2020). Medical Device Development Process, and Associated Risks and Legislative Aspects-Systematic Review. Front Public Health. 8:308
12. Mostert JM, Rietbergen DDD, Pereira Arias-Bouda L, Grootjans W (2021). Fracture risk assessment using Artificial Intelligence and Dual Energy X-ray Absorptiometry. This thesis, Chapter 2.
13. Pearson D, Horton B, Green DJ, Hosking DJ, Goodby A, Steel SA (2006). Vertebral morphometry by DXA: a comparison of supine lateral and decubitus lateral densitometers. J Clin Densitom. 9(3):295-301. https://doi.org/10.1016/j.jocd.2006.03.011
14. Bazzocchi A, Spinnato P, Fuzzi F, Diano D, Morselli-Labate AM, Sassi C, Salizzoni E, Battista G, Guglielmi G (2012). Vertebral fracture assessment by new dual-energy X-ray absorptiometry. Bone 50(4):836-841. https://doi.org/10.1016/j.bone.2012.01.018
15. Van Dort MJ, Romme EAPM, Smeenk FWJM, Geusens PPPM, Wouters EFM, van den Bergh JP (2018). Diagnosis of vertebral deformities on chest CT and DXA compared to routine lateral thoracic spine X-ray. Osteoporos Int. 29(6):1285-1293. https://doi.org/10.1007/s00198-018-4412-1
16. Mostert JM, Romeijn SR, Dibbets-Schneider P, Götz C, DiFranco MD, Dimai HP, Grootjans W (2021). Inter-operator agreement of vertebral fracture assessment with dual-energy x-ray absorptiometry equipment. This thesis, Chapter 3.
17. Mostert JM, Romeijn SR, Dibbets-Schneider P, Götz C, DiFranco MD, Grootjans W (2021). Initial validation of an Artificial Intelligence tool for automated vertebral fracture assessment. This thesis, Chapter 4.
18. Boland GWL, Guimaraes AS, Mueller PR (2009). The radiologist's conundrum: benefits and costs of increasing CT capacity and utilization. Eur Radiol. 19:9–12 .
19. McDonald RJ, Schwartz KM, Eckel LJ, Diehn FE, Hunt CH, Bartholmai BJ, Erickson BJ, Kallmes DF (2015). The effects of changes in utilization and technological advancements of cross-sectional imaging on radiologist workload. Acad Radiol. 22(9):1191-1198.

Acknowledgements

"Above all, don't fear difficult moments. The best comes from them."
- Rita Levi-Montalcini

This thesis is the result of a year of hard work, in which I have been supported by many great people. There are some people that I specifically want to thank for all their support in the past year.

First and foremost, I would like to express my appreciation to my direct supervisors Willem, Daphne, and Lenka. Thank you for all the discussions, ideas, challenges, feedback, and most importantly, the opportunity to do this project, the responsibility and freedom you gave me, and the chance to learn so many things here at the LUMC. I am also very grateful to my independent graduation committee members, Jaap and Lieselot.

I'd also like to thank everyone else involved with this project. Stephan for your help throughout the year and your feedback on this thesis, and Floris for the fruitful biweekly discussions. Petra, the initiator of this whole research project, for all your help and participation in the studies. Michael and Johan as well, for participating and for annotating many VFA's. I know you don't like doing it, so let's hope you will soon be supported by GECKO. Richard, for evaluating the radiographs – and for all the memes. And of course everyone else at the nuclear medicine department who I kept bothering for KKBtjes.

Many thanks to Christoph, Matthew, and the rest of the ImageBiopsy Lab team for your never-ending enthusiasm and for the great collaboration. Without you, this project would not have been possible. Thank you Anne, for your help with the aesthetics of this thesis.

I also should not forget my research colleagues for all their emotional support. Pim, working together was somehow always productive, despite the cats on my keyboard. Timo, thanks for your feedback, the very useful notes-holder, and the fries on Fridays. And of course Wyanne, Maaike, Fleur, Dennis, and Jeroen for the laughs and the online coffee breaks, they were a very welcome distraction in a time of working from home.

And finally, I would like to thank my friends and family, and especially mom and dad, for supporting me throughout this year. And of course Fenna, my biggest pillar of support. It wasn't an easy year, but you were always there for me when I needed it. I am incredibly proud of you.

Marijn Mostert
Leiden, 2021

Appendices

Appendix A – WCO-IOF-ESCEO abstract

Abstract accepted for presentation at the World Congress on Osteoporosis, Osteoarthritis and Musculoskeletal Diseases 2021

VERTEBRAL FRACTURE ASSESSMENT USING DUAL-ENERGY X-RAY ABSORPTIOMETRY EQUIPMENT HAS MODERATE INTEROBSERVER AGREEMENT AND AFFECTS READER FATIGUE

J. M. Mostert[1], P. Dibbets-Schneider[1], S. R. Romeijn[1], C. Götz[2], M. Difranco[2], H. P. Dimai[3], W. Grootjans[1]

[1]Leiden University Medical Center, Department of Radiology, Leiden, The Netherlands; [2] ImageBiopsy Lab, Vienna, Austria; [3]Medical University of Graz, Division of Endocrinology and Diabetology, Graz, Austria

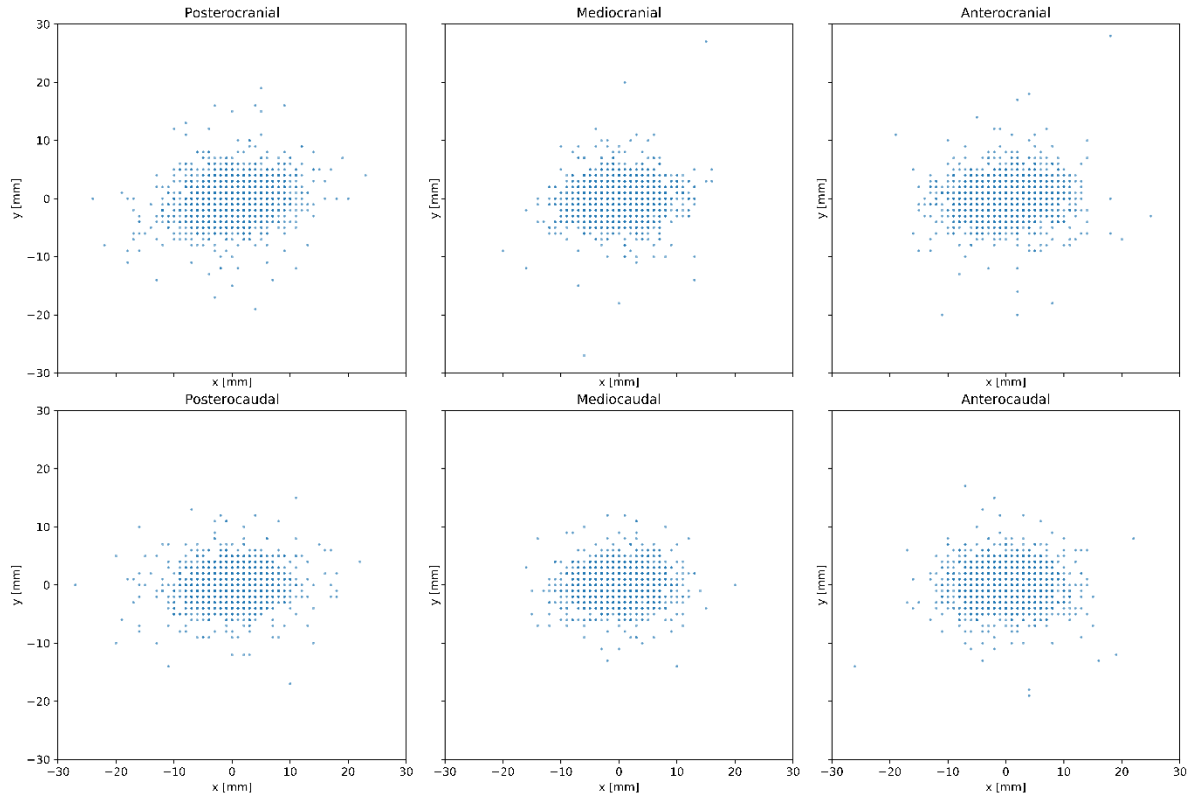
Purpose To investigate the time and effort needed to perform vertebral morphometry, as well as interobserver agreement for identification of vertebral fractures on Vertebral Fracture Assessment (VFA) images, and to evaluate the potential benefit for automated VFA.

Methods Ninety-six images were retrospectively selected and three radiographers independently performed semi-automatic 6-point morphometry in reading sessions of 6 images. Fractures were identified and graded using the Genant classification. Time needed to annotate each image was recorded and reader fatigue was assessed using a modified Simulator Sickness Questionnaire (SSQ). Interobserver agreement was assessed per-patient and per-vertebra for detecting fractures of all grades (grade 1-3) and for grade 2 and 3 fractures using the kappa statistic. Variability in measured vertebral height was evaluated using the intraclass correlation coefficient (ICC).

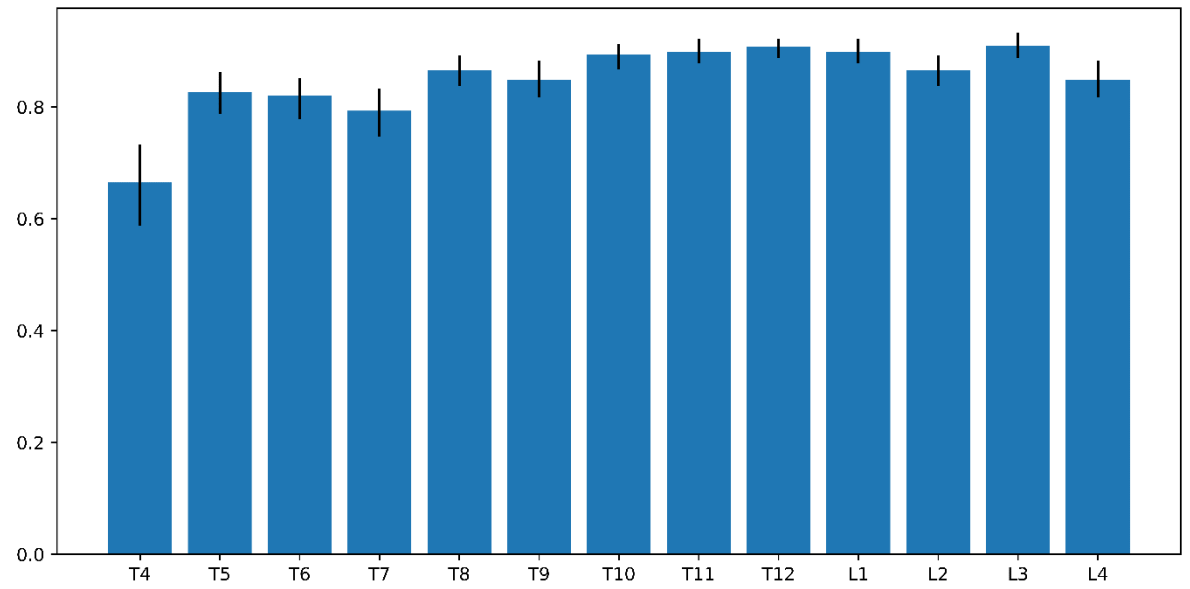
Results Per-patient agreement was 0.59 for grade 1-3 fracture detection, and 0.65 for grade 2-3 only. Agreement for per-vertebra fracture classification was 0.92. Vertebral height measurements showed a mean absolute difference from the average across radiographers of 1.38 mm (95% CI: 1.36 - 1.41), with an ICC of 0.96. Time needed to annotate VFA images ranged between 91 and 540 seconds, with a mean annotation time of 259 s. Mean SSQ scores were significantly lower at the start of a reading session (1.29; 95% CI: 0.81 - 1.77) compared to the end of a session (3.25; 95% CI: 2.60 - 3.90; $p < 0.001$).

Conclusion Although excellent ICC for vertebral height measurement was achieved, agreement for detection of patients with vertebral fractures was only moderate. In addition, vertebral morphometry requires substantial time investment and significantly affects reader fatigue. There is a potential benefit for automation tools for detection of vertebral fractures on VFA, both in improving interobserver agreement and in decreasing reading time and burden on readers.

Appendix B – Supplementary Figures to Chapter 3



Supplementary Figure A. Spread around the average landmark location across readers for manual vertebral morphometry on VFA, split per intended landmark location on the vertebra. Landmarks are translated so that the average location is at the origin, and scatter points correspond to individual annotations by one of the radiographers.



Supplementary Figure B. Intraclass Correlation for manual measurement of vertebral heights on Vertebral Fracture Assessment images, split per vertebral level. Error bars indicate 95% Confidence Intervals.

Appendix C - Simulator Sickness Questionnaire

The simulator sickness questionnaire was originally proposed by Kennedy et al. in 1993 for quantifying symptoms of users of aviation simulators. The original questionnaire involved symptoms in three clusters: nausea, disorientation, and oculomotor. In our modified SSQ scoring we only kept the oculomotor symptoms.

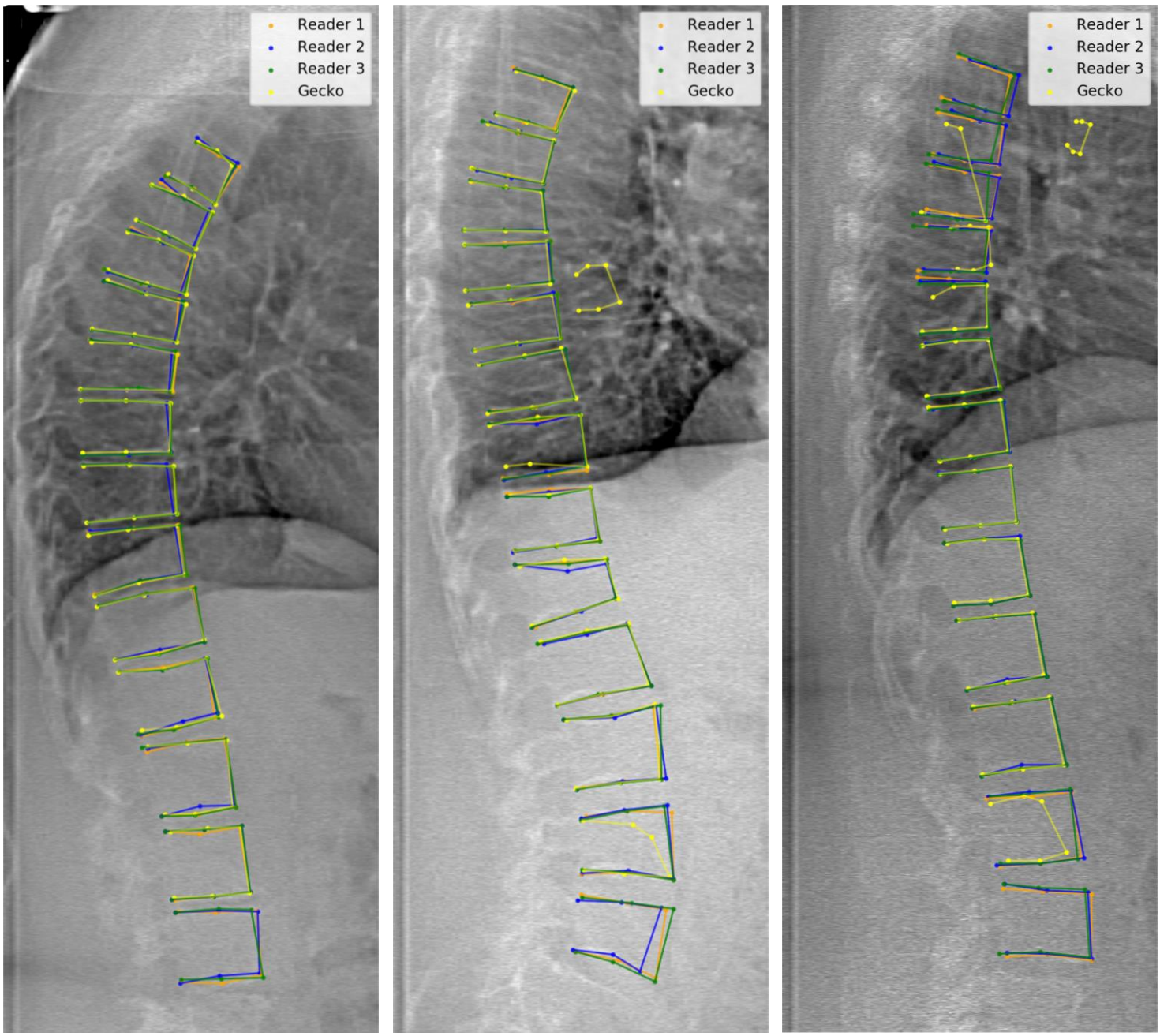
Symptom	Weight
General discomfort	1
Fatigue	1
Headache	1
Eyestrain	1
Difficulty focusing	1
Difficulty concentrating	1
Blurred vision	1

Symptoms are scored 0, 1, 2, or 3.
The total oculomotor SSQ score is the sum of the weighted scores.

Robert S. Kennedy , Norman E. Lane , Kevin S. Berbaum & Michael G. Lilienthal (1993) Simulator Sickness Questionnaire: An Enhanced Method for Quantifying Simulator Sickness, The International Journal of Aviation Psychology, 3:3, 203-220, DOI:10.1207/s15327108ijap0303_3

Appendix D – Supplementary Figures to Chapter 4

Good *Moderate* *Poor*



Supplementary Figure C. Example VFA images with good (left), moderate (middle) or poor (right) automatic annotations made by the Gecko algorithm (shown in yellow). Manual corrections by the three radiographers are also shown in orange, blue, and green.