Inter-observer agreement of vertebral fracture assessment with dual-energy x-ray absorptiometry equipment

Mostert, J.M.; Romeijn, S.R.; Dibbets-Schneider, P.; Rietbergen, D.D.D.; Arias-Bouda, L.P.M.; Gotz, C.; … ; Grootjans, W.

Archives of Osteoporosis (2022) 17:4
https://doi.org/10.1007/s11657-021-01046-w

ORIGINAL ARTICLE

Inter-observer agreement of vertebral fracture assessment with dual-energy x-ray absorptiometry equipment

Jacob M. Mostert[1] · Stephan R. Romeijn[1] · Petra Dibbets-Schneider[1] · Daphne D. D. Rietbergen[1] · Lenka M. Pereira Arias-Bouda[1,2] · Christoph Götz[3] · Matthew D. DiFranco[3] · Hans Peter Dimai[4] · Willem Grootjans[1]

Received: 12 July 2021 / Accepted: 30 November 2021 / Published online: 10 December 2021

Abstract

Purpose To investigate the time and effort needed to perform vertebral morphometry, as well as inter-observer agreement for identification of vertebral fractures on vertebral fracture assessment (VFA) images.

Methods Ninety-six images were retrospectively selected, and three radiographers independently performed semi-automatic 6-point morphometry. 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). Inter-observer agreement was assessed per-patient and per-vertebra for detecting fractures of all grades (grades 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 grades 1–3 fracture detection, and 0.65 for grades 2–3 only. Agreement for per-vertebra fracture classification was 0.92. Vertebral height measurements had an ICC of 0.96. Time needed to annotate VFA images ranged between 91 and 540 s, 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 Agreement for detection of patients with vertebral fractures was only moderate, and vertebral morphometry requires substantial time investment. This indicates that there is a potential benefit for automating VFA, both in improving inter-observer agreement and in decreasing reading time and burden on readers.

Keywords Vertebral fracture assessment · Dual-energy x-ray absorptiometry · Vertebral morphometry · Inter-observer agreement · Osteoporosis

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." Vertebrae are the most common site for osteoporotic fractures, and vertebral fractures have a major impact on patients' quality of life due to 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

✉ Willem Grootjans
w.grootjans@lumc.nl

[1] Department of Radiology, Leiden University Medical Center, Leiden, Netherlands

[2] Department of Radiology, Alrijne Hospital, Leiderdorp, Netherlands

[3] ImageBiopsy Lab, Vienna, Austria

[4] Department of Internal Medicine, Division of Endocrinology and Diabetology, Medical University Graz, Graz, Austria

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-observer variability for identification of vertebral fractures on VFA images.

Methods

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 July 2019 and 31 March 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 all patients who underwent VFA in this timeframe, a group of 96 patients was randomly selected while stratifying for age and presence of 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.

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 normal BMD, osteopenia, or osteoporosis 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.

Assessment of vertebral fractures

VFA images were independently analyzed by three clinical radiographers using Hologic Physician Viewer software (Version 7.3, Hologic, Bedford, MA, USA). In our center, these radiographers routinely perform VFA image acquisition and vertebral morphometry in clinical practice as part of our FLS. Radiographers had different levels of experience with VFA (*R1* 20 years, *R2* 10 years, and *R3* 5 years) and were blinded to each other's assessments. The study workflow is schematically depicted in Fig. 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 to determine the crush ratio.

After determining the height ratios, a grade (1–3) is assigned to the fracture, as defined by Genant et al., to

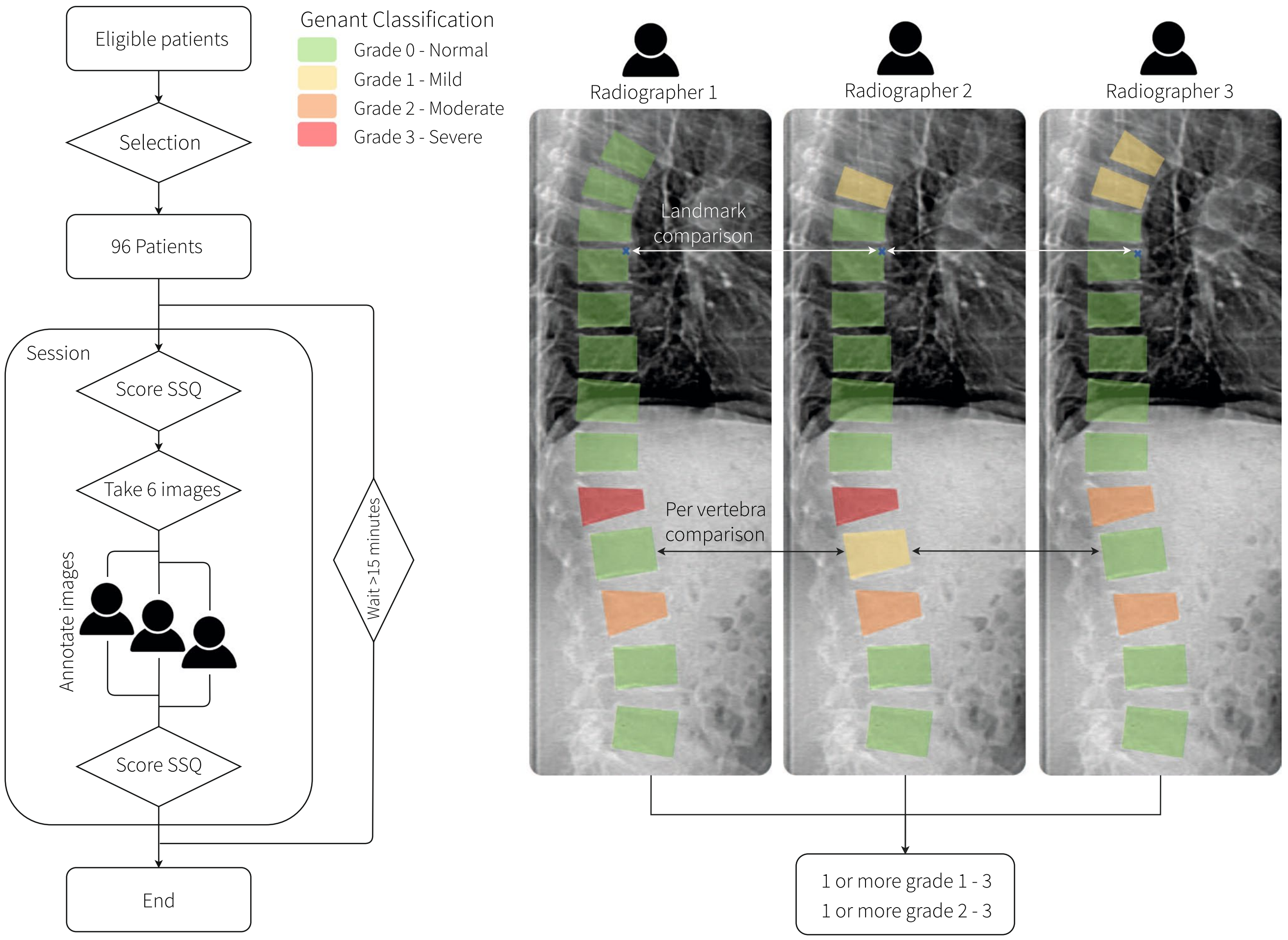

Fig. 1 Schematic description of the VFA annotation workflow in our study. Three radiographers annotated the same 96 images in sessions of 6 images, with at least 15 min between sessions. Reader fatigue was assessed by SSQ at the start and end of each annotation session. Annotation consisted of 6-point morphometry by semi-automatic landmark placement on vertebrae from T4 to L4, after which Genant classifications were automatically calculated. SSQ = Simulator Sickness Questionnaire

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). In the Genant semi-quantitative method, readers estimate vertebral heights and do not perform quantitative morphometry on all vertebrae. In contrast, we used a method in which quantitative morphometry is performed on all vertebrae in the image, after which vertebrae are graded using the same Genant grading system. Only vertebrae that the radiographers deemed not evaluable due to insufficient image quality or image artefacts were not annotated and 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 using the Genant semi-quantitative classification method by an experienced radiologist who was not involved in VFA annotation. With this method, vertebral height deviations were taken into account, as well as morphological characteristics including endplate fracture, cortical buckling, lack of endplate parallelism, and loss of vertical continuity.

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. The time needed to annotate each image was recorded. Between different annotation sessions, a break of at least 15 min was planned. At the start and end of each session, the readers filled in a modified oculomotor Simulator Sickness Questionnaire (SSQ) to assess reader fatigue [18]. The readers were asked to score the presence of 7 common symptoms (general discomfort, fatigue, headache, eyestrain, difficulty focusing, difficulty concentrating, blurred vision) on a 5-point scale; the overall SSQ score was given as the sum of these scores.

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 inter-observer agreement, the kappa statistic was calculated for detecting fractures of all grades (grades 1–3) and for moderate and severe (grades 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.

In addition, inter-observer 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 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].

In 57 patients (60%), spinal radiographs were available and 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.

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 (32%) 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 labeling 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.

Fracture classification

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 grades 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 Randolph's kappa score of 0.59. When including only grades 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 Fig. 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 (R2 & R3: 0.62) than between the most experienced radiographer and the less experienced radiographers (R1 & R2: 0.53; R1 & R3: 0.57).

For 57 patients, a spinal radiograph was available. Of these, 28 patients (49%) had at least one vertebral fracture grades 1–3, and 16 patients (28%) had one or more vertebral fractures of grades 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 scores between VFA and spinal radiography were 0.51 for R1, 0.58 for R2, and 0.54 for R3 for fractures regardless of severity. When including only grade 2–3 fractures, kappa scores were 0.52, 0.57, and 0.60, respectively. When considering fracture detection on spinal radiographs as the ground truth, the radiographers

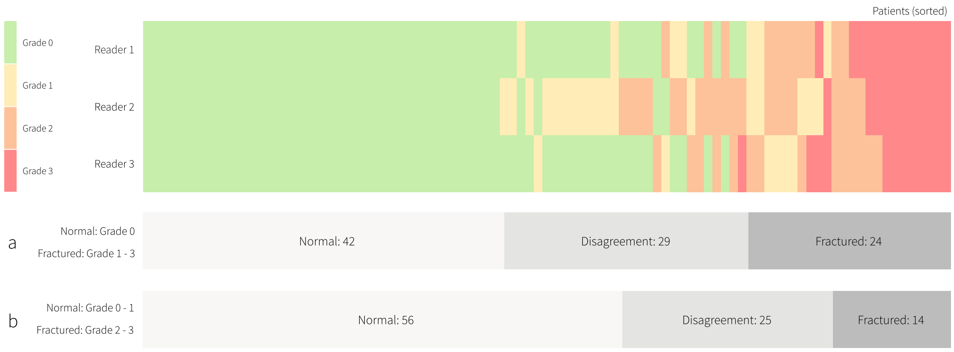

Fig. 2 Highest detected Genant classification per patient as determined by the three radiographers. Patients are classified as fractured or not fractured, either including grade 1 (**a**) or excluding grade 1 (**b**)

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.

Per vertebra analysis

Of the 1235 vertebrae included, 121 (9.8%) 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 Fig. 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 (grades 2–3) by another radiographer. For 6 patients, this affected the highest-grade vertebra and would have affected fracture diagnosis.

Height measurement and landmark placement

Landmark placement was analyzed for all 1127 evaluable 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

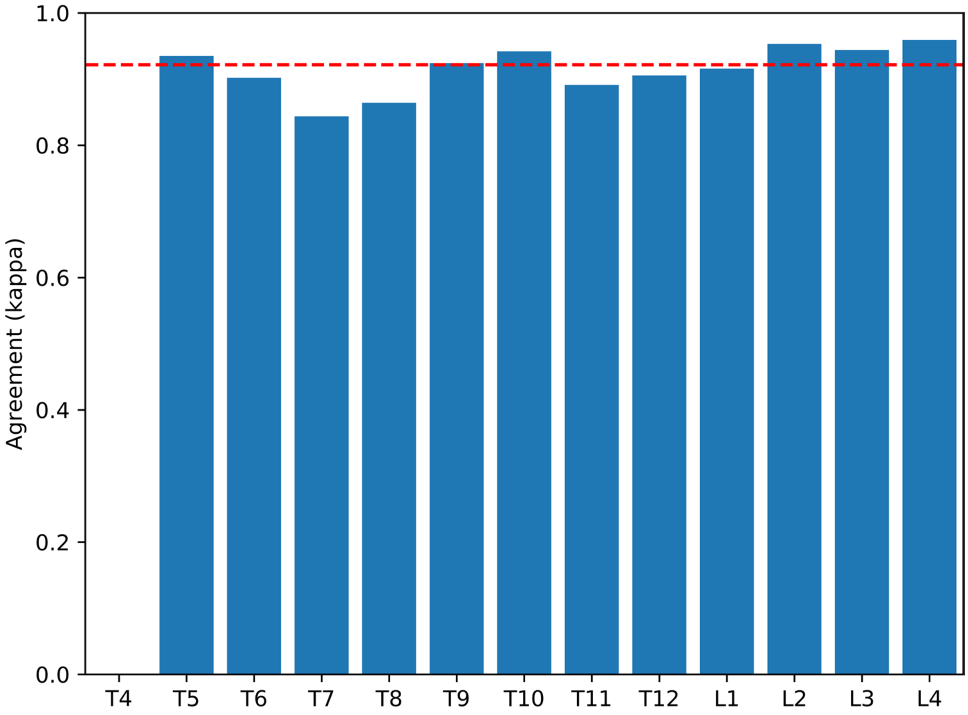

Fig. 3 Agreement (Randolph's kappa) per vertebral level for vertebral fracture severity classification on VFA. The red dashed line indicates overall agreement level. Agreement for T4 could not be determined since all included T4 vertebrae were considered grade 0 by all radiographers

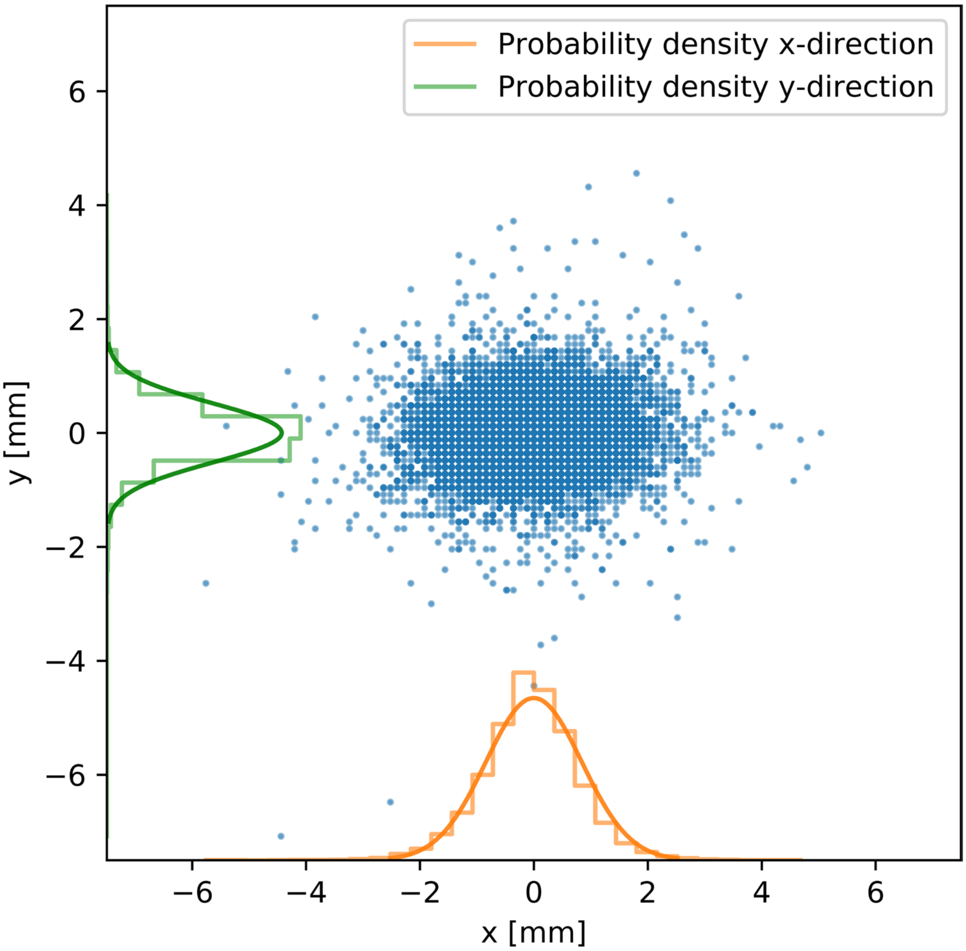

Fig. 4 Spread around the average location for each landmark. Landmarks are translated so that the average location is at the origin, and scatter points correspond to individual annotations by one of the radiographers

of 0.81 mm (95% CI: 0.80–0.82). The spread around the mean for each landmark is shown in Fig. 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 (grades 2–3) or not fractured (grades 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.

Reading time and reader discomfort

The time needed to annotate the VFA images ranged between 91 and 540 s, 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 reading session (3.25; 95% CI: 2.60–3.90; $p < 0.001$). SSQ scores per reader for each reading session are shown in Fig. 5.

Discussion

Inter-observer agreement for vertebral fracture assessment

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 fair 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 a missed fracture diagnosis, which has happened in 6 cases in this study. Agreement was slightly lower when grade 1

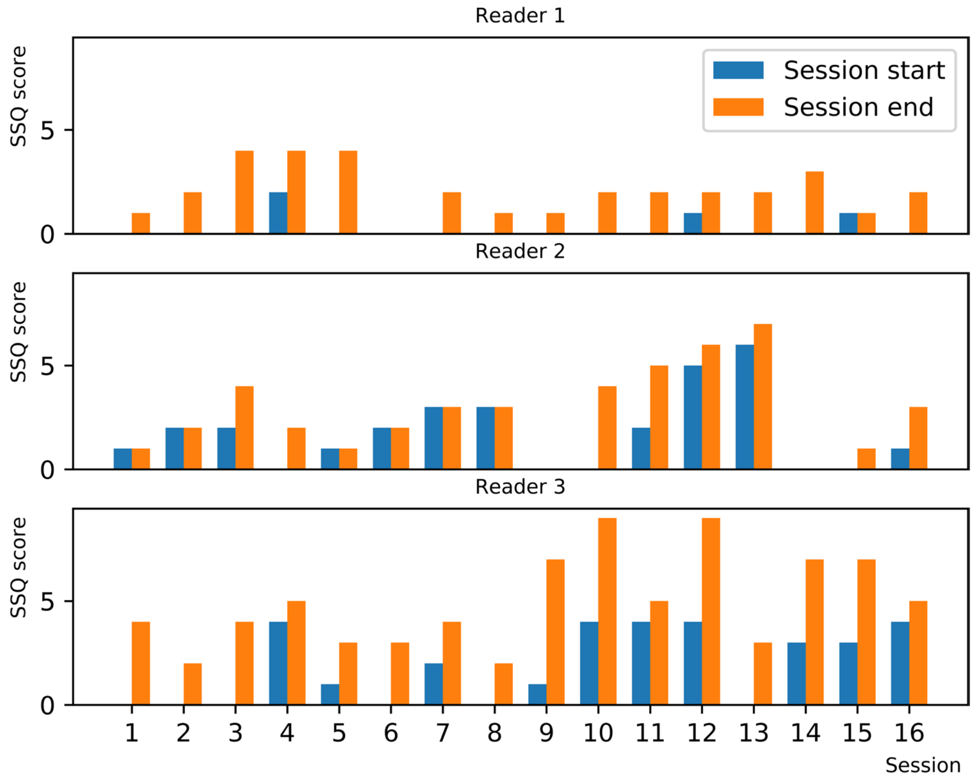

Fig. 5 Reader discomfort per radiographer for annotating VFA images. Scores are given for the start and end of each reading session of six images. SSQ: Simulator Sickness Questionnaire

fractures were also included, which seems to be induced by readers' difficulty in differentiating grade 1 deformities from normal vertebrae.

Despite the moderate inter-observer agreement, variability of landmark placement for vertebral morphometry on VFA images is very small. On average, a landmark was placed 0.81 mm from its mean location across annotators. The average height of thoracolumbar vertebrae ranges between 22 and 36 mm, so a deviation of 0.81 mm corresponds with 2.3% to 3.7% of vertebral height [21]. Although this is only a small fraction of the total vertebral height, deviations close to classification cut-off values can lead to unwanted differences in fracture classification. For patients where all radiographers agreed on the fracture status, the mean absolute distance was smaller than for patients where there was disagreement. Although statistically significant, absolute differences were very small, further indicating that with quantitative morphometry using Genant classification, multiple readers placing landmarks very close together can still yield different classification results. As such, fracture classification with vertebral morphometry is highly sensitive to small variations in landmark placement.

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.

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 s per vertebra. This is very similar to our findings, as an average annotation time of 259 s per patient gives 22.1 s per vertebra, when accounting for non-evaluable vertebrae.

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.

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.

For this study, we selected a patient cohort with a vertebral fracture prevalence much higher than would be expected in a random clinical population. Kappa statistics can be affected by the prevalence of an attribute when the proportion of agreement on positive classifications differs from the proportion of agreement on negative classifications. Although Randolph's free-marginal multirater kappa coefficient is not affected by this, extrapolation of our results to clinical populations should be done with caution.

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 min 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 Fig. 5, it seems that baseline SSQ scores increase for a number of consecutive sessions. This cumulative fatigue may indicate that 15 min 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.

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. In addition, vertebral morphometry is time-consuming and has a significant effect on reader fatigue. Especially in FLS with high patient numbers, there could be a potential benefit for automation tools for detection of vertebral fractures on VFA. However, automation tools should focus on clinically relevant outcome measures such as agreement with conventional radiographic imaging.

Acknowledgements The authors would like to thank Michael Bruijns and Johan van Brecht for their enthusiasm and time annotating VFA images and Richard Raghoo, MD, for evaluating the spinal radiographs.

Data availability The dataset from this study is not publicly available due to data protection regulations.

Code availability Statistical analyses were performed using open-source python modules Scikit-learn 0.20.3, Statsmodels 0.10.1, and Pingouin 0.3.8.

Declarations

Ethics approval All procedures performed in studies involving human participants were in accordance with the 1964 Helsinki declaration and its later amendments and the Medical Ethics Committee of the Leiden University Medical Center waived the need to acquire informed consent (registration number G20.032).

Consent to participate Not applicable.

Consent for publication Not applicable.

Competing interests Authors CG and MD are employees of Image-Biopsy Lab.

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
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) 50:2264–2269. 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 22(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 7:221–227. 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):557–563. 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

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.