Check for updates

RESEARCH ARTICLE

JBMR®

Towards Improved Identification of Vertebral Fractures in Routine Computed Tomography (CT) Scans: Development and External Validation of a Machine Learning Algorithm

Joeri Nicolaes,[1,2] Michael Kriegbaum Skjødt,[3,4] Steven Raeymaeckers,[5] Christopher Dyer Smith,[4] Bo Abrahamsen,[3,4,6] Thomas Fuerst,[7] Marc Debois,[2] Dirk Vandermeulen,[1] and Cesar Libanati[2]

[1]Department of Electrical Engineering (ESAT), Center for Processing Speech and Images, KU Leuven, Leuven, Belgium
[2]UCB Pharma, Brussels, Belgium
[3]Department of Medicine, Hospital of Holbæk, Holbæk, Denmark
[4]OPEN–Open Patient Data Explorative Network, Department of Clinical Research, University of Southern Denmark and Odense University Hospital, Odense, Denmark
[5]Department of Radiology, Universitair Ziekenhuis Brussel, Brussels, Belgium
[6]NDORMS, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, Oxford University Hospitals, Oxford, UK
[7]Clario, Princeton, NJ, USA

ABSTRACT

Vertebral fractures (VFs) are the hallmark of osteoporosis, being one of the most frequent types of fragility fracture and an early sign of the disease. They are associated with significant morbidity and mortality. VFs are incidentally found in one out of five imaging studies, however, more than half of the VFs are not identified nor reported in patient computed tomography (CT) scans. Our study aimed to develop a machine learning algorithm to identify VFs in abdominal/chest CT scans and evaluate its performance. We acquired two independent data sets of routine abdominal/chest CT scans of patients aged 50 years or older: a training set of 1011 scans from a non-interventional, prospective proof-of-concept study at the Universitair Ziekenhuis (UZ) Brussel and a validation set of 2000 subjects from an observational cohort study at the Hospital of Holbæk. Both data sets were externally reevaluated to identify reference standard VF readings using the Genant semiquantitative (SQ) grading. Four independent models have been trained in a cross-validation experiment using the training set and an ensemble of four models has been applied to the external validation set. The validation set contained 15.3% scans with one or more VF (SQ2-3), whereas 663 of 24,930 evaluable vertebrae (2.7%) were fractured (SQ2-3) as per reference standard readings. Comparison of the ensemble model with the reference standard readings in identifying subjects with one or more moderate or severe VF resulted in an area under the receiver operating characteristic curve (AUROC) of 0.88 (95% confidence interval [CI], 0.85–0.90), accuracy of 0.92 (95% CI, 0.91–0.93), kappa of 0.72 (95% CI, 0.67–0.76), sensitivity of 0.81 (95% CI, 0.76–0.85), and specificity of 0.95 (95% CI, 0.93–0.96). We demonstrated that a machine learning algorithm trained for VF detection achieved strong performance on an external validation set. It has the potential to support healthcare professionals with the early identification of VFs and prevention of future fragility fractures. © 2023 UCB S.A. and The Authors. *Journal of Bone and Mineral Research* published by Wiley Periodicals LLC on behalf of American Society for Bone and Mineral Research (ASBMR).

KEY WORDS: OSTEOPOROSIS; SPINAL FRACTURES; MACHINE LEARNING; TOMOGRAPHY; X-RAY COMPUTED; HUMANS

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.
Received in original form February 17, 2023; revised form September 6, 2023; accepted September 17, 2023.
Address correspondence to: Joeri Nicolaes, MSc, Department of Electrical Engineering (ESAT), Center for Processing Speech and Images, KU Leuven, Belgium.
E-mail: joeri.nicolaes@ucb.com
Additional Supporting Information may be found in the online version of this article.

Journal of Bone and Mineral Research, Vol. 00, No. 00, Month 2023, pp 1–11.

DOI: 10.1002/jbmr.4916
© 2023 UCB S.A. and The Authors. *Journal of Bone and Mineral Research* published by Wiley Periodicals LLC on behalf of American Society for Bone and Mineral Research (ASBMR).

Introduction

Osteoporosis affects approximately 200 million people globally, resulting in 9 million fragility fractures each year.[1,2] Vertebral fractures (VFs) due to osteoporosis are common with one occurring every 22 seconds worldwide in individuals aged 50 years or older. Patients suffering from VFs are exposed to a higher risk for subsequent fractures, notably hip fractures,[3] and face an increased risk of death compared to patients without VFs.[4] Accurate identification of VFs may, therefore, offer a means to flag patients at high risk for future debilitating fracture.[5] However, it is estimated that up to one of three VFs come to clinical attention.[6] VFs are often asymptomatic or mildly symptomatic and radiological studies are most often obtained for other clinical purposes. The workload of radiologists increased by double digits over the last decades driven by an increase in cross-sectional imaging studies in routine and emergency care, two-thirds of which were computed tomography (CT) exams.[7-9] Despite this increase in available images, opportunistic identification of VFs by radiologists in imaging studies visualizing the spine is lagging and many VFs go undetected or unreported.[10,11]

Osteoporosis imaging is used to quantitatively assess bone quality and diagnose prevalent fragility fractures such as VFs. Radiologists can identify VFs in radiographs, dual-energy X-ray absorptiometry (DXA) images, and sagittal reformations of CT images[12] by applying different reading standards that can be categorized as qualitative, quantitative, and semiquantitative (SQ) assessments. The qualitative approach relies on the reader's expertise to visually assess morphologic features and is a subjective method with poor interobserver agreement. Quantitative methods rely on morphometric features, are objective and reproducible, yet they lack specificity for VFs. Genant's SQ assessment combines morphometric (eg, shape) with morphologic (eg, endplate deformity) features[13] and is recommended by most societies such as the International Society for Clinical Densitometry (ISCD), International Osteoporosis Foundation (IOF), and European Society of Musculoskeletal Radiology (ESSR).[12] The SQ method is commonly applied in research studies as a gold standard. It is considered more objective and reproducible than a qualitative approach, but can be difficult to apply.[12,14] Several studies have shown that interreader and intrareader variability can be significant across modalities and for different VF grades.[15,16] Finally, recent work showed that the debate on the most appropriate reading standard is still very much ongoing.[17]

Machine learning algorithms can support VF identification by opportunistically evaluating CT scans of abdomen or chest for suspected VFs that can be confirmed by a radiologist and reported to a healthcare professional.[18] Computer-aided diagnosis (CAD) methods for VF detection are applied to 2D and 3D modalities yet most exclusively leverage 2D information (ie, sagittal reformations in the case of CT).[19] Modeling approaches range from segmentation of vertebral bodies followed by height measurements to deep learning methods automatically scoring an image as containing VFs or not.[20-25] Previous diagnostic performance studies reported single-center validation results on sample sizes of a few hundred subjects.[24,26] One study reported subject-level fracture detection results on 1700 subjects but applied an adjudication procedure unblinded to CAD readings, which may have inflated their performance results.[27] Over the last two decades, machine learning methods, such as deep learning based methods, have been successfully applied to detect and segment objects in images.[28] However, deep learning methods for medical image analysis still face several challenges, such as data availability, generalizability, interpretability, and uncertainty quantification for which research is still ongoing.[29]

VFs are the hallmark of osteoporosis and are associated with a marked increase in future fracture risk.[3] Yet the majority go undetected and the opportunistic identification of VFs in routine imaging exams is lagging. Therefore, the objective of this study was to develop a new automated algorithm capable of identifying VFs in abdomen/chest CT scans and to evaluate its performance against blinded reference standard readings in an external validation set.

Materials and Methods

In this diagnostic accuracy study, we acquired two independent data sets of abdominal/chest CT scans of subjects aged 50 years or older, performed for various indications. The first data set of 1011 routine CT scans from a non-interventional, prospective proof-of-concept study at UZ Brussel (Belgium) was used as a training set to develop a machine algorithm for the automated detection of VFs. The second data set of 2000 CT scans from an observational cohort study performed at Holbæk Hospital (Denmark) was used as an external validation set to evaluate the algorithm's performance. Both data sets were evaluated to identify prevalent VFs and establish the reference standard readings for every vertebra visible in these scans. A validation sample of 204 subjects with VFs, which would be surpassed assuming 15% prevalence in the validation set, was estimated to be sufficient to measure a sensitivity of 80% requiring the lower 95% confidence limit to be >70% with 95% probability.[30]

We developed an automated VF detection algorithm and assessed its performance on two outcomes: (i) a subject-level binary outcome for the presence of one or more VFs in the CT scan, and (ii) a vertebral-level binary outcome for the presence of a VF for every vertebra visible in the CT scan. We apply two binarization schemes to the SQ grades: (i) outcome VF_{SQ123} with normal in the "no VF" and mild, moderate, and severe VFs in the "VF" category, and (ii) outcome VF_{SQ23} with normal and mild VFs in the "no VF" and moderate and severe VFs in the "VF" category. Although the former outcome captures every SQ grade defined by the Genant method, the latter focuses on the clinically most important fractures and is therefore considered the primary outcome of this study.

The reporting of this study followed the standards for the reporting of diagnostic accuracy studies (STARD) 2015 guidelines.[31]

Study population

Training set

The training set contained 1011 routine CT scans from a non-interventional, prospective proof-of-concept study, performed at UZ Brussel between January and August 2019. The Ethical Committee at UZ Brussel approved this study (B.U.N. 143201732477) and informed consent has been acquired from all subjects enrolled in this study.

CT scans of the abdomen (potentially including the pelvis) and chest were identified from routine care at UZ Brussel by one

15234681, 0, Downloaded from https://asbmr.onlinelibrary.wiley.com/doi/10.1002/jbmr.4916, Wiley Online Library on [17/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

board-certified radiologist (SR) using the following selection criteria: (i) include subjects aged 50 years or older at the time of the scan, (ii) construct a random sample of male / female subjects, (iii) construct a balanced sample of abdominal and chest CT scans, (iv) construct a balanced sample of "VF"/"no VF" scans (presence of any VF was identified through visual inspection by SR), and (v) construct a cumulative sample of ≥10 VFs of Genant SQ grades mild, moderate, and severe from thoracic or lumbar vertebrae. All the available CT studies were extracted as Digital Imaging and Communications in Medicine (DICOM) images from the radiology database and pseudonymized after extraction.

The 1011 scans contained between 2 and 20 diverse CT studies (different reconstructions, reformations and occasionally different CT exams acquired during the same visit). JN manually reviewed all CT scans in Osirix MD Viewer to include one CT study for each subject in the training set using the following criteria: (i) exclude CT scans with a slice thickness of >3 mm, (ii) maximize the cumulative number of VFs for SQ grades mild, moderate, and severe for every vertebral level T_1 to L_5, and (iii) maximize variability of scans across different manufacturer models, exam types (abdomen, abdomen with hip, chest, full spine), convolution reconstruction kernels, slice thickness, in-plane pixel spacing and signal-to-noise ratio (qualitatively assessed by JN). This resulted in a total of 921 CT scans, excluding 90 CT scans belonging to duplicate subjects. Vertebral centroids and levels were manually annotated by JN using MeVisLab.[32]

Validation set

The CT scans of the validation set were retrospectively acquired in the context of an observational cohort study, approved by the Danish Patient Safety Authority (3-3013-2687/1), Statistics Denmark (707480), and covered by the Danish Data Protection Agency approval for Region Zealand healthcare research (REG-101-2018). Ethics committee approval was not required for the validation set.

An external validation set of 2000 abdominal/chest CT scans performed at Holbæk Hospital from January 1, 2010 onward was extracted from the radiology database and pseudonymized after extraction. The CT scans were retrospectively acquired in the context of an observational cohort study, approved by the Danish Patient Safety Authority (3-3013-2687/1), Statistics Denmark (707480), and covered by the Danish Data Protection Agency approval for Region Zealand healthcare research (REG-101-2018). The study included male/female subjects aged 50 years or older at the time of the scan. The data obtained in the CT scan reevaluation was linked—on an individual level—to the Danish national registers. From these, information on demographics, medical history, and use of pharmaceutical drugs were obtained. Further details on the eligibility criteria, methods, sample size calculations and baseline characteristics of the validation scans can be found in Skjødt and colleagues.[33] The validation set data flow diagram is shown in Fig. 1. CT scans identified by the machine learning algorithm as "not readable" and/or not having any registry data were excluded from the analyses reported here.

Reference standard reading

Reference standard readings were established in both training and validation sets by evaluating the CT scans for prevalent VFs in a two-step process blinded to clinical information. First, a trained medical doctor (CL) triaged the scans in three categories (certain VF, potential VF, and no VF). Second, vertebral-level reference standard readings were produced by highly experienced radiologists (Clario, USA) using the semiquantitative (SQ) Genant classification.[13] Dr. Harry Genant adapted the SQ method from radiographs to CT scans and supervised the standardization sessions with the reading team prior to the commencement of reading activity. A single radiologist from the reading team evaluated every visible vertebra in the scans categorized with certain VF or potential VF, together with a 5% subset of scans without VFs. Radiologists were blinded to the scan selection process and the triage category. Reader variability assessments were performed for a subset (10%) of all scans. In this article, we refer to normal vertebrae as SQ0, mild VF as SQ1, moderate VF as SQ2, and severe VF as SQ3. Figure S1 illustrates the reference standard reading for every SQ grade on exemplary vertebrae extracted from our training set.

Development of VF detection algorithm

We developed an automated VF detection algorithm that comprised of two components: (i) a VF detection model that determines the SQ grade (ie, SQ0-3), and (ii) a vertebra identification model that localizes each vertebra present in the scan. This algorithm processes an abdominal/chest CT scan and outputs an estimated SQ grade for every vertebra identified in the scan (Fig. 2A-E). The VF detection results can be visualized as heat maps overlaid on top of the original CT scan.

We developed a VF detection Convolutional Neural Network (CNN) model by extending on our previous work that was limited to detecting binary, subject-level VFs and was trained on a smaller dataset of 90 CT scans.[34] CT scans were resampled to $1 \times 1 \times 1$ mm^3, and intensities were clipped between -1024 and 2000 Hounsfield units and Z-score normalized as preprocessing steps. CT scans with corrupt DICOM series (eg, due to missing slices or erroneous DICOM headers) were not readable by the algorithm. The machine learning algorithm automatically outputs the "not readable" signal. The VF detection model ingests 3D patches of size $114 \times 114 \times 114$ mm^3 extracted from the preprocessed CT image and processes each patch at normal and subsampled resolution (Fig. S2). The features learned at both scales are concatenated and further postprocessed to output a class probability for every SQ grade. We model the output classes as independent variables estimated using a linear model of the shared features and trained the model with a binary cross-entropy loss. This design allows the CNN to learn discriminating features for each SQ grade and to output mixed belief between grades if needed. The training set for the VF detection CNN model was constructed by merging the VF reference standards readings with the vertebra centroids annotations on vertebral level for the 921 CT scans. JN manually reviewed the CT scans with merged readings in MeVisLab to identify vertebra label mismatches or missing centroids. CT scans with such mismatches or misses, together with scans that revealed image quality issues during this manual review were excluded from the training set. We used the vertebra localization model developed by Payer and colleagues[35] to estimate centroids and levels for every vertebra present in the CT image. This model achieved state-of-the-art identification and localization results of approximately 90% in consecutive years on the Verse challenge.[36]

The 4D VF detection probability maps were aggregated to vertebral level by averaging the voxel probabilities in a 3D cube of size S around every estimated centroid. This resulted in a belief score vector s_i for every vertebra i found in the CT scan corresponding to the model's belief for every SQ grade (Fig. 2E). The

15234681, 0, Downloaded from https://asbmr.onlinelibrary.wiley.com/doi/10.1002/jbmr.4916, Wiley Online Library on [17/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

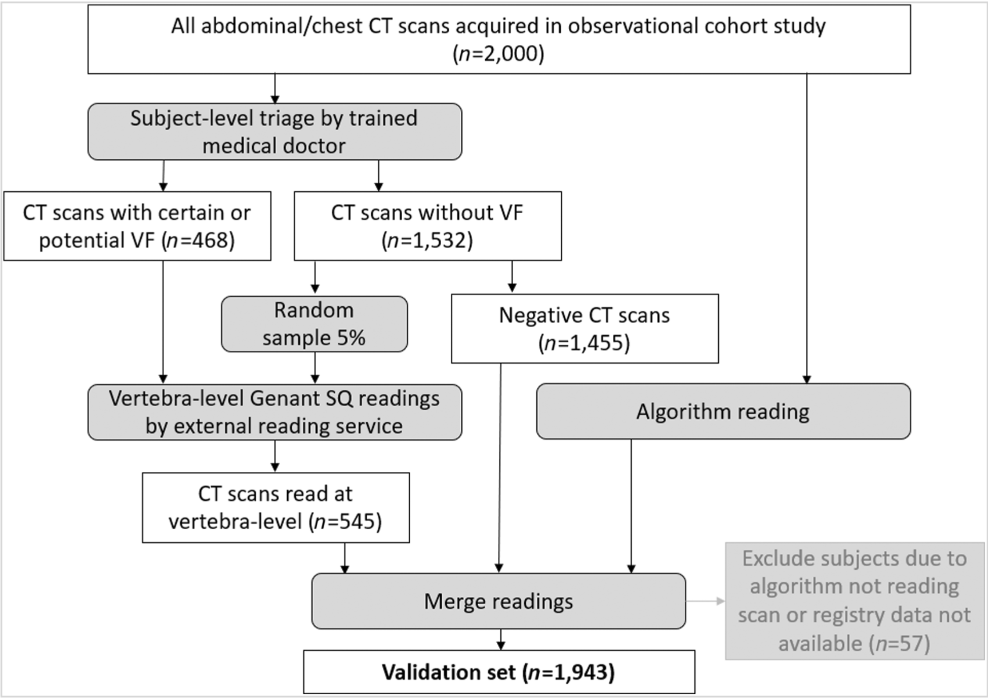
All abdominal/chest CT scans acquired in observational cohort study (n=2,000)
Subject-level triage by trained medical doctor
CT scans with certain or potential VF (n=468)
CT scans without VF (n=1,532)
Random sample 5%
Negative CT scans (n=1,455)
Vertebra-level Genant SQ readings by external reading service
Algorithm reading
CT scans read at vertebra-level (n=545)
Merge readings
Exclude subjects due to algorithm not reading scan or registry data not available (n=57)
Validation set (n=1,943)

Fig. 1. Data flow diagram of the validation set: 2000 abdominal/chest CT scans were collected in an observational cohort study, performed at Holbaek hospital from January 2010 onward. The gray boxes depict a processing activity performed on the data. CT = computed tomography; SQ = semiquantitative grade; VF = vertebral fracture.

scores s_i were discretized to an estimated SQ grade $v_i \in$ {SQ0, SQ1, SQ2, SQ3} by combining the beliefs for the different SQ grades using a belief function: $v_i = f(s_i, t_1, t_2)$. This function encodes that the estimated SQ grade v_i is the output with the highest belief if this belief score is above a given minimum belief threshold t_1. If the highest and second highest belief scores are close within a given mixed belief threshold t_2, the model outputs a mixed belief from the set {normal/mild, mild/moderate, moderate/severe} to encode that both SQ grades are as likely for the model. The binary vertebral-level and binary subject-level outcomes were derived in a straightforward manner from the vertebral-level, categorical outcomes v_i.

We conducted a four-fold cross-validation (CV) experiment to develop the VF detection CNN model by applying a stratified split of the training set. We stratified all CT scans into 24 groups representing all the combinations of the worst fracture grade (SQ0, SQ1, SQ2, SQ3), exam type (abdominal, chest, full spine) and gender (M, F) using the scikit learn package.[37] In each fold, three splits have been used as training set (of which 15% of the scans were held out to determine the optimal hyperparameters, ie, best epoch, minimum belief threshold t_1 and mixed belief threshold t_2 of each fold model) and one split was used as a CV test set. Models were trained for 40 epochs and the best epoch and minimum belief threshold t_1 were determined for each fold model using the mean F1-scores of the validation samples. The mixed belief threshold t_2 was set in a principled manner to allow no more than 10% of the vertebrae to be detected with mixed belief. All other hyperparameters were defaulted to state-of-the-art settings for all fold models (e.g., RMSprop optimizer, batch normalization enabled, L_1–L_2 regularization and dropout enabled during training). Each fold model was trained on one NVidia GTX 1080 Ti GPU card and a custom-developed Python 3 software package that leverages Tensorflow[38] version 2.3 and SimpleITK[39] version 1.2. We applied a cube size $S = 10$ for the vertebral-level aggregation. We used the open-source docker image with weights trained on the Verse2020 challenge data for the vertebra localization model. We used the latest model available at https://hub.docker.com/r/christianpayer/verse20.

We applied an ensemble of four independently trained models by averaging the vertebral level scores s_i for each grade across all models. This approach is similar to asking four experts to independently read a CT scan for VFs and weighing each expert's opinion equally in a consensus review. The algorithm and reference standard readings were blinded to each other.

Statistical analyses

Baseline characteristics are presented using median and interquartile range (IQR) and counts and proportions for continuous and categorical variables respectively. Groups were compared by two-tailed Student *t* tests (training set, Table 1) and median test (validation set, Table 2) for continuous data and χ^2 tests for categorical data. A significance threshold of $p < 0.05$ was applied.

The validation set was used to evaluate the performance of the algorithm in identifying VFs at the subject and vertebral level by the area under the receiver operating characteristic curve (AUROC), accuracy, sensitivity, specificity, positive and negative predictive values (PPV/NPV), and Cohen's kappa. We used the interpretations defined in[40] for Cohen's kappa. The accelerated bootstrapping method with bias correction applying 1000

15234681, 0, Downloaded from https://asbmr.onlinelibrary.wiley.com/doi/10.1002/jbmr.4916, Wiley Online Library on [17/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

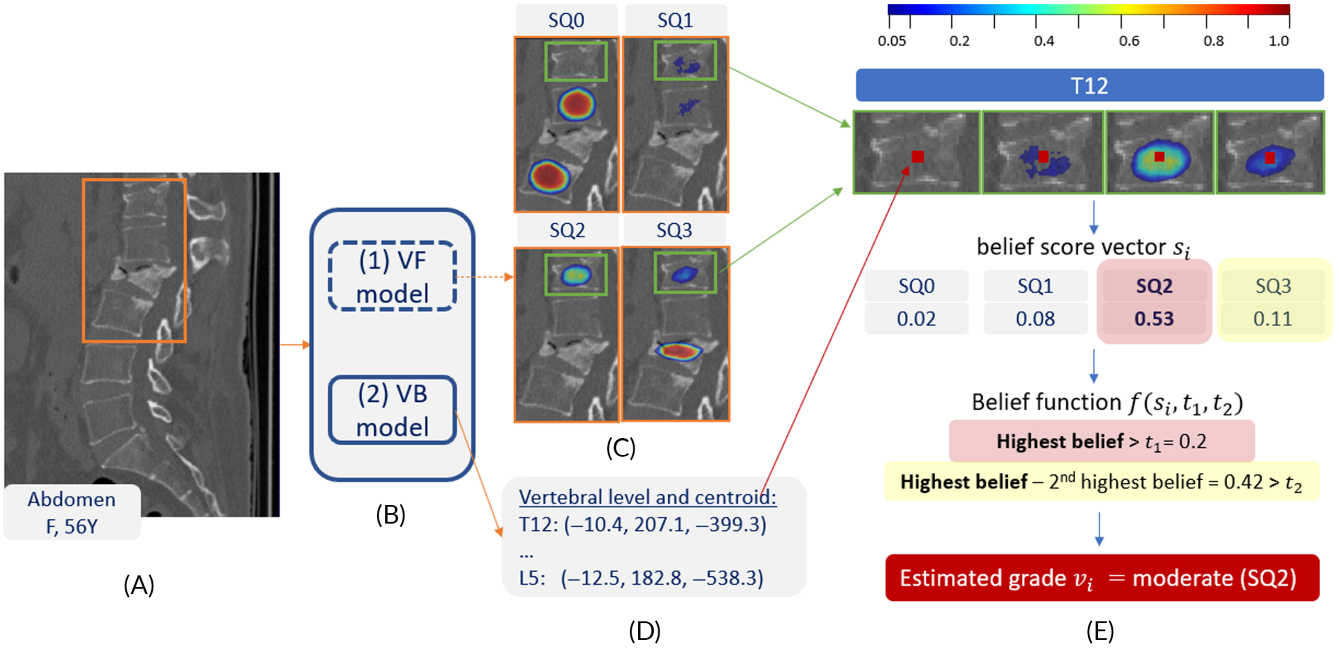
SQ0
SQ1
SQ2
SQ3
0.05
0.2
0.4
0.6
0.8
1.0
T12
belief score vector s_i
SQ0
SQ1
SQ2
SQ3
0.02
0.08
0.53
0.11
Belief function $f(s_i, t_1, t_2)$
Highest belief > t_1= 0.2
Highest belief – 2nd highest belief = 0.42 > t_2
Estimated grade v_i = moderate (SQ2)
(1) VF model
(2) VB model
Abdomen
F, 56Y
Vertebral level and centroid:
T12: (–10.4, 207.1, –399.3)
...
L5: (–12.5, 182.8, –538.3)
(A)
(B)
(C)
(D)
(E)

Fig. 2. Automated VF detection algorithm overview: schematic to illustrate how the algorithm estimates the SQ grade for the top T_{12} vertebra in an abdominal scan from our training set. A 3D patch is extracted from the CT scan (*A*) and processed by two independent models for (1) identifying the SQ grade and (2) localizing the vertebrae present in the scan respectively (*B*). The VF model outputs a VF detection heat map image for every SQ grade representing the algorithm's confidence that a given SQ grade is present at voxel level (*C*). The VB model outputs a list of vertebral levels identified with their centroid coordinates (*D*). For every vertebra *i* identified by the VB model, a postprocessing step computes a score vector s_i by averaging the confidence scores in the VF detection heatmap around the vertebra centroid (illustrated with the red point) and defines the estimated grade from the belief function as depicted. In this example, the estimated grade for T_{12} is SQ2 because the belief score vector s_i has the highest belief score for SQ2 and this SQ2 belief score is $>t_1$ and $>t_2+$ second highest SQ3 belief score. Each image displays one mid-sagittal slice extracted from the CT scan, after thresholding intensities between -1024 and 2000 Hounsfield units. Images in *C* and *E* additionally show the VF detection heat maps overlaid on top of the sagittal slice, using heat maps colors from 0.05 (blue) to 1.0 (red) as depicted in *E*. (*A*) Abdominal CT scan from training set; (*B*) VF detection algorithm models; (*C*) VF detection heat map output images; (*D*) VB identification model output list; (*E*) postprocessing the VF and VB outputs illustrated for the top T_{12} vertebra. CT = computed tomography; SQ = semiquantitative grade; VB = vertebra; VF = vertebral fracture.

repetitions was used to construct the 95% CIs. We computed subject-level and vertebral-level results for the primary outcome VF_{SQ23} (negative = SQ0 or SQ1, positive = SQ2 or SQ3) and the secondary outcome VF_{SQ123} (negative = SQ0, positive = SQ1, SQ2 or SQ3). We conducted a subgroup analysis on the primary outcome VF_{SQ23} at the subject level (age, gender) and at the vertebral level (gender, CT exam type). Sensitivity analyses were performed, firstly, to exclude subjects treated with an osteoporosis medication (OM) at any time before baseline, and subjects with a diagnosis code for any malignancies, Paget's disease, and/or monogenetic osteoporosis at baseline, and secondly, to exclude subjects treated with corticosteroids within the year prior to baseline. Finally, we analyzed the identification performance of the vertebra localization model by reporting the ratio between the number of scans for which the first and last vertebra level were identified identically and the total number of scans.

Differences in vertebra labeling by the reference standard reading and algorithm resulted in a subset of the vertebrae being graded by one yet marked as "not visible" by the other. For the vertebral-level analysis, we regrouped the "not visible" vertebrae together with SQ0-1 vertebrae. Vertebrae identified as "not visible" by both reference standard reading and algorithm were excluded from the vertebral-level analysis. The performance analysis was executed twice on the validation set, using algorithm version 1 readings (performed in 2021) and using algorithm version 2 readings (2022–2023). None of the validation cases used in current study were used to update the algorithm from version 1 to 2. The VF detection model was not re-trained for algorithm 2; only the belief function has changed from version 1 to 2. We described and presented only the results from the latest algorithm version 2 in this manuscript. We compared the mixed belief outputs (normal/mild, mild/moderate, moderate/severe) forgivingly by accepting an agreement if the reference standard reading was present in the model output (eg, if the algorithm detected normal/mild, a correct detection was recorded if the reference standard reading was normal *or* mild and a miss if the reference standard reading was moderate or severe). Statistical analyses comparing the reference standard readings with the algorithm outputs in the validation set were performed using Stata version 16 and 17 (StataCorp, College Station, TX, USA).

Results

Study data

The 921 CT studies belonging to unique subjects have been linked with the VF readings and vertebral centroid annotations

15234681, 0, Downloaded from https://asbmr.onlinelibrary.wiley.com/doi/10.1002/jbmr.4916, Wiley Online Library on [17/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 1. Baseline Characteristics of the Training Set Stratified in "VF_{SQ123} (SQ grade >0)"/"no VF" Groups

Characteristic	No VF (SQ 0)	VF_{SQ123}	*p*
Total, *N* (%)	299 (45%)	367 (55%)	
Subject demographics			
Gender, female, *n* (%)	134 (45%)	221 (60%)	<0.001
Age, years, median (IQR)	70 (61–77)	78 (69–84)	<0.001
CT exam			
Visible vertebrae; median (IQR)	10 (8–14)	12 (8–15)	<0.001
Abdomen exams, *n* (%)	11 (4%)	156 (43%)	<0.001
Chest exams, *n* (%)	210 (70%)	82 (22%)	<0.001
Full spine exams, *n* (%)	77 (26%)	124 (34%)	0.03

Note: The comparisons between both groups are statistically significant on all presented characteristics ($p < 0.05$).

Abbreviations: IQR = interquartile range; SQ = semiquantitative grade; VF = vertebral fracture.

Table 2. Baseline Characteristics of the Validation Set Stratified in "VF_{SQ123} (SQ grade >0)"/"no VF" Groups

Characteristic	No VF (SQ 0)	VF_{SQ123}	*p*-value
Total, *N* (%)	1536 (79%)	407 (21%)	
Subject demographics			
Gender, female, *n* (%)	707 (46%)	218 (54%)	0.007
Age, years; median (IQR)	68 (61–76)	74 (67–80)	<0.001
Country of origin, Denmark, *n* (%)	1473 (96%)	396 (97%)	0.19
Medical history			
CCI score; median (IQR)	1 (0–2)	2 (0–3)	<0.001
Major osteoporotic fracture, *n* (%)[a]	141 (9%)	103 (25%)	<0.001
Anti-osteoporosis medication, *n* (%)[b]	73 (5%)	106 (26%)	<0.001
Glucocorticoid therapy, *n* (%)[c]	230 (15%)	89 (22%)	<0.001

Note: The comparisons between both groups are statistically significant on all presented characteristics ($p < 0.05$) except for the country of origin.

Abbreviations: CCI = Charlson comorbidity index; IQR = interquartile range; $_{SQ123}$ = mild, moderate, or severe VF; SQ = semiquantitative grade; VF = vertebral fracture.

[a] Major osteoporotic fracture is defined as hip, non-cervical vertebral, humerus and distal forearm fracture.

[b] At any time prior to baseline.

[c] In the year prior to scan.

to generate a training set of 666 samples, excluding 255 training samples because of vertebral level mismatches (50%), missing centroids (30%), and image quality issues (20%). This training set contained 292 (44%) chest (T_4 is visible), 167 (25%) abdominal (L_2 is visible), 201 (30%) full spine (T_4 and L_2 are visible) and six other (neither T_4 nor L_2 are visible) CT scans. These scans originated from four different scanners (Philips ICT 256, GE Revolution CT, GE Discovery CT750 HD, Siemens Definition AS40), had a median slice thickness of 1 mm (IQR: 0.9–1.25 mm, minimum: 0.625 mm, maximum: 3 mm), a median in-plane pixel-spacing of 0.68 mm (IQR: 0.58–0.74 mm, minimum: 0.24 mm, maximum: 1.22 mm), three different peak kilovoltage (kVp) outputs of 100 (45%), 120 (47%), and 140 (8%) and 23 different convolution kernels (12 of which are present more than 10 times). The training set contains a total of 367 scans (55%) with one or more VFs, representing 7537 vertebrae of which 915 (12%) are fractured (SQ grade >0). Figure 3 shows the number of VFs stratified to SQ grade in the training set across all vertebrae with a bimodal distribution for the proportion of visible vertebrae that are fractured as reported by others.[24] The expected amount of at least 10 VFs of every SQ grade per vertebral level has been reached for all but vertebrae T_1–T_4, T_{10}, and L_5. We found that scans with ≥1 VF are predominantly female subjects who are on average older than subjects without VFs and that chest exams are predominantly present in the "no VF" groups (Table 1). We performed a reader variability study on a subset of 62 scans and a total of 766 vertebrae in a challenging set containing more than 80% of CT scans with one or more VFs. We found at the subject level a Cohen's kappa of 0.61 (95% CI, 0.38–0.84) and 0.76 (95% CI, 0.54–0.98) for the primary outcome VF_{SQ23} and secondary outcome VF_{SQ123}, respectively, and at the vertebral level a Cohen's kappa of 0.69 (95% CI, 0.62–0.76) and 0.80 (95% CI, 0.76–0.85) for the primary outcome VF_{SQ23} and secondary outcome VF_{SQ123}, respectively.

In the validation set, the scan of some subjects was not readable by the algorithm and some subjects had no registry data available. These subjects ($n = 57$) were excluded. Of the remaining 1943 scans, 297 (15.3%) had one or more VF_{SQ23} and 407 (20.9%) had one or more VF_{SQ123}, whereas 663 of 24,930 vertebrae (2.7%) were fractured with SQ grade 2–3 and 1066 of 24,930 vertebrae (4.3%) were fractured with SQ grade 1–3. Figure 4 shows the number of VFs stratified to SQ grade in the validation set across all vertebral levels with a bimodal distribution peaking at T_7–T_8 and T_{12}–L_1. The CT scans in the validation set have been acquired on one scanner model (Philips Brilliance 64) and were all secondary DICOM images resampled to a slice thickness of 3 mm. The CT scans had a median in-plane pixel-spacing of 0.76 mm (IQR: 0.68–0.88 mm, minimum: 0.18 mm, maximum: 1.61 mm). Ninety-eight percent (98%) of the CT scans were acquired using a kVp of 120 (29 and 11 scans were acquired using 100 and 140 kVp, respectively). 56% of all CT scans were thorax exams. Table 2 shows the baseline characteristics of the "VF_{SQ123}" and the "no VF" cohorts in the validation set. The median Charlson comorbidity index (CCI) score and the proportion of subjects with a major osteoporotic fracture (defined as hip, non-cervical vertebral, humerus, and distal forearm fracture) were higher in the VF_{SQ123} cohort, and a larger proportion of the VF_{SQ123} subjects received anti-osteoporosis medication (any time prior to scan) and glucocorticoid therapy (in the year before scan), all *p* values <0.001 (Table 2). We performed a reader variability study on 50 scans, representing a total of 594 vertebrae. The validation set subsample used for the reader variability assessment was representative for the validation set, both for vertebra-level prevalence of SQ grades 0–3 and for subject-level VF prevalence. We found at the subject level a Cohen's kappa of 0.72 (95% CI, 0.52–0.93) and 0.77 (95% CI, 0.52–1.00) for the primary outcome VF_{SQ23} and secondary outcome VF_{SQ123}, respectively, and at the vertebral level a Cohen's kappa of 0.83 (95% CI, 0.77–0.90) and 0.78 (95% CI, 0.72–0.85) for the primary outcome VF_{SQ23} and secondary outcome VF_{SQ123}, respectively.

15234681, 0, Downloaded from https://asbmr.onlinelibrary.wiley.com/doi/10.1002/jbmr.4916, Wiley Online Library on [17/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

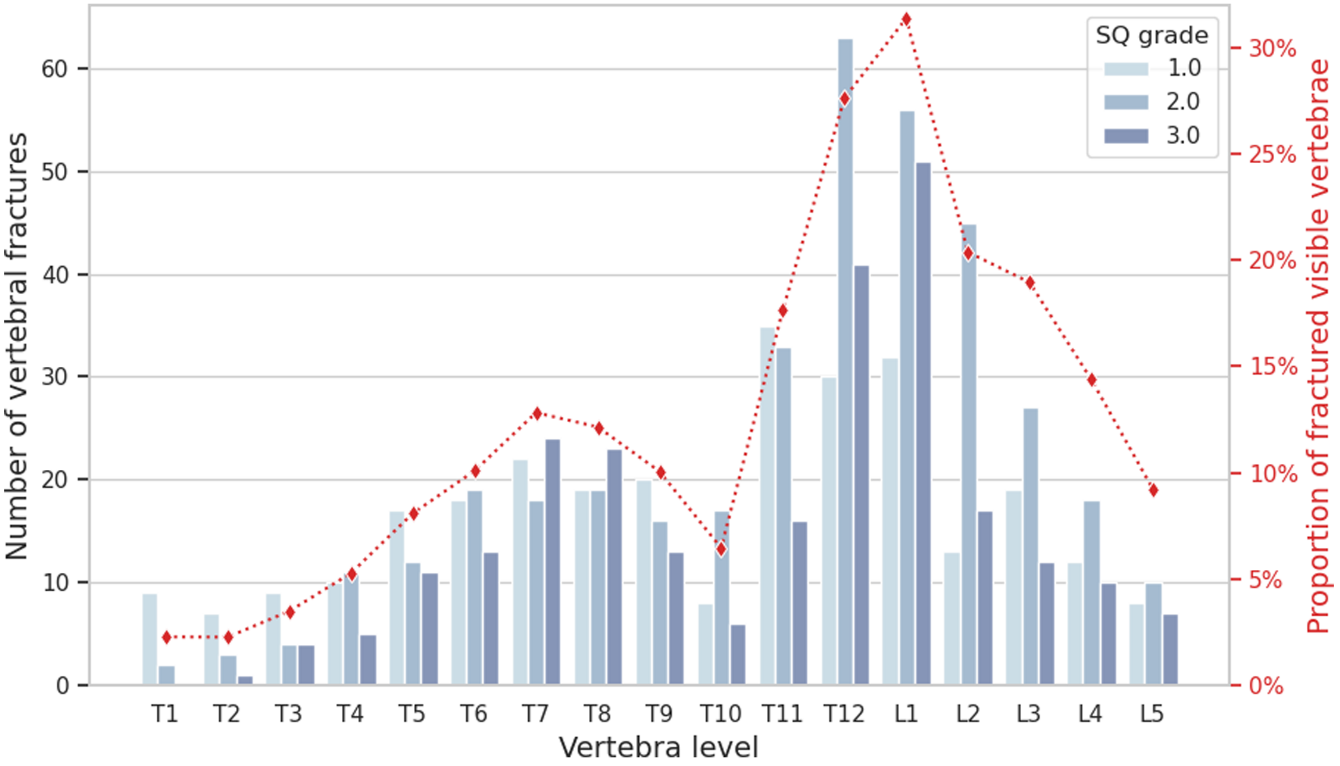

Fig. 3. Training set: number of fractured vertebrae per level and SQ grade. The horizontal axis depicts the vertebrae in standard anatomic fashion, starting from T_1 or the first thoracic vertebra and ending with L_5 or the fifth lumbar vertebra. The total number of VFs stratified according to SQ grade (left axis, vertical bars) and proportion of visible vertebrae that are fractured (red color, right axis, diamond points) are shown. SQ = semiquantitative grade; VB = vertebral body; VF = vertebral fracture.

Algorithm development

We trained four VF detection CNN models using the architecture in Fig. S2, each containing a total of 112,517 trainable parameters. Most computations during training were central processing unit (CPU)-bound (eg, preprocessing, data augmentation, and patch sampling) resulting in runs of more than 100 hours to train each model on a single graphics processing unit (GPU). We found that model training demonstrated sound learning curves with monotonically decreasing training and cross-validation losses. The cross-validation analysis showed that the vertebral-level results using the belief function $v_i = f(s_i, t_1, t_2)$ were stable for different values of the minimum belief threshold t_1 for all fold models. Although any choice between 0.0 and 0.25 would be reasonable from the cross-validation analysis, we set t_1 to 0.2 for all fold models as a conservative choice that would avoid noisy detections, which we specifically found for edge vertebrae in our cross-validation analysis. The mixed belief threshold t_2 was set to 0.20, 0.18, 0.13, and 0.13 for the fold1, fold2, fold3, and fold4 models, respectively, from the cross-validation analysis.

Diagnostic performance in validation set

The metrics for the evaluation of the diagnostic performance of the VF detection algorithm versus reference standard readings in the validation set are shown in Table 3A,B for outcomes SQ23 and SQ123, respectively. Confusion matrices for all outcomes can be found in Fig. S3. The SQ23 subject-level performance in differentiating normal/mild from moderate/severe VFs showed an AUROC of 0.876 (95% CI, 0.852–0.898), a Cohen's kappa of 0.72 (95% CI, 0.67–0.76), an accuracy of 92.4% (1795/1943), a sensitivity of 80.8% (240/297), a specificity of 94.5% (1555/1646), a PPV of 72.5% (240/331) and an NPV of 96.5% (1555/1612). The vertebral-level performance for identifying VF_{SQ23} had an AUROC of 0.763 (95% CI, 0.745–0.783), a Cohen's kappa of 0.58 (95% CI, 0.55–0.62), an accuracy of 98.1% (24,444/24,930), a sensitivity of 53.2% (353/663), a specificity of 99.3% (24,091/24,267), a PPV of 66.7% (353/529), and an NPV of 98.7% (24,091/24,401). The machine learning algorithm detected a total of 638 (2.6%) vertebrae with mixed belief outputs (normal/mild, mild/moderate, moderate/severe), of which 128 vertebrae were correctly matched with reference standard readings based on the second highest belief.

We found in the subgroup analysis at the subject level, an accuracy of 92.8% (858/925) and 92.0% (937/1018) and a Cohen's kappa of 0.77 and 0.64 for women and men, respectively, and an accuracy of 94.0% (914/972) and 90.7% (881/971) and a Cohen's kappa of 0.69 and 0.73 for the younger (50–69 years) and older (70+ years) age groups, respectively (Table 4A). We found in the subgroup analysis at the vertebral level, an accuracy of 97.5% (11,465/11,758) and 98.5% (12,979/13,172) and a Cohen's kappa of 0.60 and 0.55 for female and male vertebrae, respectively, and an accuracy of 98.2% (16,503/16,810) and 97.8% (7941/8120) and a Cohen's kappa of 0.54 and 0.64 for the thoracic and lumbar vertebrae, respectively (Table 4B). We found an accuracy of 92.8% (1065/1148) and 92.6% (1504/1624) and a Cohen's kappa of 0.66 and 0.72 in the first (excluding subjects

15234681, 0, Downloaded from https://asbmr.onlinelibrary.wiley.com/doi/10.1002/jbmr.4916, Wiley Online Library on [17/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

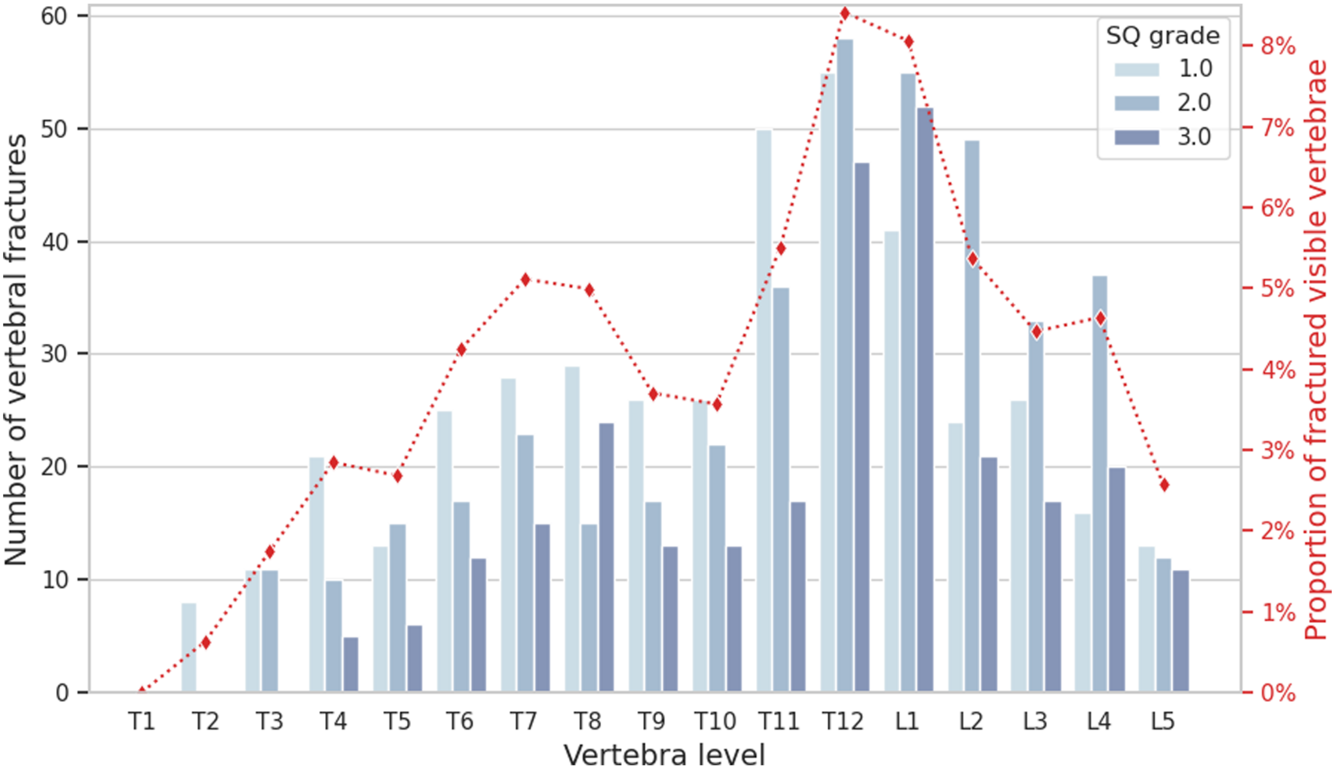

Fig. 4. Validation set: number of fractured vertebrae per level and SQ grade including all 2000 subjects (25,391 vertebrae visualized in CT scan). The horizontal axis depicts the vertebrae in standard anatomic fashion, starting from T_1 or the first thoracic vertebra and ending with L_5 or the fifth lumbar vertebra. The total number of VFs stratified according to SQ grade (left axis, vertical bars) and proportion of visible vertebrae that are fractured (red color, right axis, diamond points) are shown. VF numbers (and proportions) have not been reported if $N < 5$ or if discretized to maintain confidentiality. Figure adapted from[33] with permission. SQ = semiquantitative grade; VF = vertebral fracture.

Table 3. Diagnostic Performance of VF Detection Algorithm Versus Reference Standard Readings in the Validation Set

Metric	Subject level	Vertebral level
(*A*) SQ23 (normal/mild versus moderate/severe)		
AUROC	0.876 (0.852–0.898)	0.763 (0.745–0.783)
Accuracy	0.92 (0.91–0.93)	0.98 (0.98–0.98)
Kappa	0.72 (0.67–0.76)	0.58 (0.55–0.62)
Sensitivity	0.808 (0.762–0.851)	0.532 (0.494–0.569)
Specificity	0.945 (0.933–0.955)	0.993 (0.992–0.994)
PPV	0.725 (0.675–0.770)	0.667 (0.625–0.705)
NPV	0.965 (0.955–0.973)	0.987 (0.986–0.989)
(*B*) SQ123 (normal versus mild/moderate/severe) outcome		
AUROC	0.815 (0.790–0.837)	0.728 (0.714–0.744)
Accuracy	0.85 (0.83–0.87)	0.96 (0.96–0.97)
Kappa	0.58 (0.54–0.63)	0.52 (0.49–0.55)
Sensitivity	0.757 (0.714–0.798)	0.470 (0.438–0.499)
Specificity	0.873 (0.855–0.889)	0.987 (0.986–0.988)
PPV	0.612 (0.569–0.653)	0.616 (0.582–0.648)
NPV	0.931 (0.918–0.943)	0.977 (0.974–0.979)

Note: The metrics are stratified by outcome SQ23 (normal and mild versus grade 2–3) and SQ123 (normal versus grade 1–3), and unit of analysis (subject and vertebra). The depicted numbers are point estimates with 95% CI between parentheses, all CI have been generated using a bias-corrected and accelerated bootstrapping method (1000 iterations).

Abbreviations: AUROC = area under the receiver operating characteristic curve; CI = confidence interval; NPV = negative predictive value; PPV = positive predictive value; SQ = semiquantitative grade; VF = vertebral fracture.

15234681, 0, Downloaded from https://asbmr.onlinelibrary.wiley.com/doi/10.1002/jbmr.4916, Wiley Online Library on [17/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 4. Subgroup Analysis of VF Detection Algorithm Versus Reference Standard Readings for Outcome SQ23 (Normal or Mild Versus Moderate or Severe VF) in the Validation Set

Metric	Female	Male	50–69 years	70+ years
(*A*) Subject level				
AUROC	0.905 (0.876–0.930)	0.836 (0.790–0.869)	0.892 (0.851–0.927)	0.867 (0.838–0.893)
Accuracy	0.93 (0.91–0.94)	0.92 (0.90–0.94)	0.94 (0.92–0.95)	0.91 (0.89–0.93)
Kappa	0.77 (0.72–0.82)	0.64 (0.56–0.71)	0.69 (0.60–0.76)	0.73 (0.67–0.78)
Sensitivity	0.868 (0.812–0.909)	0.724 (0.635–0.795)	0.833 (0.736–0.902)	0.797 (0.739–0.850)
Specificity	0.941 (0.923–0.957)	0.948 (0.931–0.961)	0.951 (0.936–0.964)	0.937 (0.918–0.952)
PPV	0.774 (0.710–0.828)	0.654 (0.568–0.734)	0.636 (0.546–0.718)	0.775 (0.717–0.827)
NPV	0.969 (0.954–0.979)	0.962 (0.947–0.973)	0.982 (0.971–0.991)	0.945 (0.926–0.989)
Metric	**Female**	**Male**	**Thoracic**	**Lumbar**
(*B*) Vertebral level				
AUROC	0.774 (0.749–0.797)	0.742 (0.711–0.776)	0.748 (0.722–0.775)	0.782 (0.752–0.809)
Accuracy	0.98 (0.97–0.98)	0.99 (0.98–0.99)	0.98 (0.98–0.98)	0.98 (0.98–0.98)
Kappa	0.60 (0.56–0.64)	0.55 (0.49–0.60)	0.54 (0.49–0.58)	0.64 (0.59–0.69)
Sensitivity	0.557 (0.507–0.604)	0.490 (0.427–0.551)	0.503 (0.449–0.553)	0.570 (0.512–0.627)
Specificity	0.991 (0.989–0.992)	0.995 (0.993–0.996)	0.993 (0.991–0.994)	0.993 (0.991–0.995)
PPV	0.683 (0.636–0.734)	0.638 (0.573–0.708)	0.602 (0.544–0.658)	0.759 (0.704–0.812)
NPV	0.984 (0.982–0.986)	0.990 (0.989–0.992)	0.989 (0.987–0.990)	0.984 (0.981–0.987)

Note: We compared subgroups according to age and gender at subject-level (*A*), and gender and CT exam type at vertebral-level (*B*). The depicted numbers are point estimates with 95% CI between parentheses. All CIs have been generated using a bias-corrected and accelerated bootstrapping method (1000 iterations).

Abbreviations: AUROC = area under the receiver operating characteristic curve; CI = confidence interval. NPV = negative predictive value; PPV = positive predictive value; SQ = semiquantitative grade; VF = vertebral fracture.

treated with OM at any time before baseline or with a diagnosis code for any malignancies, Paget's disease, and/or monogenetic osteoporosis at baseline), and second (excluding subjects treated with corticosteroids within the year prior to baseline) sensitivity analysis, respectively.

The SQ123 subject-level performance in differentiating normal from grade 1–3 VFs showed an AUROC of 0.815 (95% CI, 0.790–0.837), a Cohen's kappa of 0.58 (95% CI, 0.54–0.63), an accuracy of 84.9% (1649/1943), a sensitivity of 75.7% (308/407), a specificity of 87.3% (1341/1536), a PPV of 61.2% (308/503), and an NPV of 93.1% (1341/1440). The vertebral-level performance for identifying VF$_{SQ123}$ had an AUROC of 0.728 (95% CI, 0.714–0.744), a Cohen's kappa of 0.52 (95% CI, 0.49–0.55), an accuracy of 96.5% (24,053/24,930), a sensitivity of 47.0% (501/1066), a specificity of 98.7% (23,552/23,864), a PPV of 61.6% (501/813), and an NPV of 97.7% (23,552/24,117).

The reference standard readings and the vertebra localization model agreed on the first and last identifiable vertebra level in 64% of the scans.

Discussion

In this study, we found that a machine learning algorithm trained for identifying VFs on abdominal and chest CT scans from one center demonstrated excellent diagnostic performance on 1943 validation CT scans from another center. The algorithm reached a Cohen's kappa score of 0.72, a sensitivity of 81%, and a specificity of 95% compared to reference standard readings in identifying subjects with moderate or severe VFs in the external validation set. The current clinical approach relies on radiologists identifying prevalent VFs in any CT scan that visualizes the (abdominal and/or thoracic) spine. Although only one of three VFs are diagnosed clinically,[6] leaving a substantial diagnostic gap, other studies have demonstrated that less than one of six incidental VFs on CT exams are reported by radiologists.[10,11] Hence, our machine learning algorithm compares favorably to current clinical practice in terms of sensitivity.

Vertebral-level SQ grade reference standard readings were defined by expert readers in a review process blinded to clinical information and algorithm readings. We found in the reader variability assessment a moderate agreement for the subject-level and vertebral-level primary and secondary outcomes in both the training and validation sets. The training set was constructed to contain more than 10 VFs of every SQ grade at every vertebral level (Fig. 3). Both training and validation set showed a bimodal distribution of VFs along the spine as reported by others.[24] In the validation set, we found a prevalence of 15% for moderate or severe VFs (VF$_{SQ23}$) and 21% for mild, moderate, or severe VFs (VF$_{SQ123}$) in line with the previous literature.[26,27] The VF$_{SQ123}$ cohort was older and sicker (median CCI score 2 versus 1) than the "No VF" cohort in the validation set (Table 2, *p* values <0.001).

Applied in a clinical alerting workflow where every positive case would be manually confirmed by an expert, the VF$_{SQ23}$ algorithm would detect 81% of the positive cases and reduce the number of scans to be reviewed by the radiologist by a factor of six (331 CT scans flagged positive by the algorithm out of 1943 CT scans in total). The VF detection algorithm demonstrated better agreement in identifying moderate or severe VFs (VF$_{SQ23}$) than mild, moderate, or severe VFs (Cohen's kappa score of 0.72 versus 0.58), suggesting that the algorithm performed best on the clinically most important VFs, as studies have not shown strong evidence of an association between mild VFs and low bone mineral density.[14] The lower performance on mild VFs was expected as mild VF readings exhibit higher inter-reader variability as reported[15] and are thus more ambiguous to read for both an expert reader and the algorithm. We accepted the mixed belief outputs forgivingly (Section Statistical analyses) to deal with this ambiguity, yet this biases the results upward. The vertebral-level results should be interpreted with caution because mismatches

15234681, 0, Downloaded from https://asbmr.onlinelibrary.wiley.com/doi/10.1002/jbmr.4916, Wiley Online Library on [17/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

between the vertebral levels identified by the model and reader are common and these mismatches void the vertebral-level results but importantly, they do not influence the subject-level results (ie, the reader and model detecting a moderate VF but disagreeing on its level, eg, T_{12} versus L_1, yields an agreement at the subject level but not at the vertebral level). Such vertebral-level disagreement could be resolved by a radiologist after visual inspection of the algorithm's outputs (shown in Fig. 2C) and hence, this issue would be alleviated in clinical practice. Furthermore, when the whole spine is not visible, the exact localization of a specific vertebral body may be challenging. We argue that while the vertebral-level results provide insight on the correct functioning of the algorithm, the subject-level results are the most important clinical outcomes that inform on how the algorithm can aid with the identification of patients with VFs to inform treatment decisions. We found minor differences in the performance between subgroups with the algorithm performing marginally better in women than in men, in the older (70+ years) versus younger (50–69 years) age groups, and in lumbar versus thoracic vertebrae.

Prior work studied the performance of a different VF detection algorithm in a validation set of 1700 CT scans acquired at one center.[27] The study applied an adjudication procedure that unblinded the algorithm's outputs to the readers and VFs were only read and evaluated at the subject level, not at the vertebral level. In a cohort different than that of our study, the authors reported a lower subject-level sensitivity of 65% (versus 81% in our study) and a similar specificity of 92% (versus 95% in our study). Another study constructed a balanced validation set of 500 CT scans (half with and half without VFs, VF reading executed by an automated machine learning method) involving the lumbar spine, randomly selected from the radiology database of the network of university hospitals of the Greater Paris Area.[26] They reported a sensitivity and a specificity of 94% (95% CI, 89%–98%) and 65% (95% CI, 60%–70%), respectively, for the subject-level binary SQ23 outcome.

In our study, we constructed an external validation of abdominal and chest CT scans performed more than a decade ago at a single center and containing subjects of predominantly Danish descent. Future studies should assess the algorithm's performance in contemporary CT scans performed on subjects of different ethnicity, acquired on different scanners in clinical centers across the globe. We evaluated the performance of an algorithm compared to reference standard readings, yet future work should study how the application of such algorithm impacts clinical care using two study arms, ie, one with and one without an algorithm as computer-aided support. Finally, a sufficiently powered subgroup analysis should be conducted to assess whether the algorithm performs similarly in subjects of different age, gender, and ethnicity, considering the statistical differences that exist between those subgroups.

Regardless of future work our study provides evidence that automatic assistance in the identification of VFs, which are largely omitted or ignored in current radiological practice, is feasible with sufficient accuracy and sensitivity to become a useful tool to assist overcoming the clinical workload and ultimately improve patient care.

Conclusion

In summary, we demonstrated that an automated algorithm trained for identifying VFs achieved excellent performance in an external validation cohort of abdominal and chest CT scans of Danish patients ≥50 years. Such an algorithm has the potential to bridge the known reporting gap by opportunistically screening for VFs in routine CT scans and flagging the scans that need attention to the radiologist. This in turn would be expected to improve the earlier appropriate management of patients at very high risk of future fractures.

Acknowledgments

This study was funded by UCB Pharma and Amgen Inc. The external validation set originates from a cohort study that was funded by UCB/Amgen and Region Zealand Health Scientific Research Foundation. This manuscript was reviewed by UCB Pharma and Amgen Inc. prior to submission.

Author Contributions

Joeri Nicolaes: Methodology; software; formal analysis; data curation; investigation; writing – original draft. **Michael Kriegbaum Skjødt:** Methodology; formal analysis; investigation; writing – review and editing. **Steven Raeymaeckers:** Investigation; data curation; writing – review and editing. **Christopher Dyer Smith:** Validation; writing – review and editing. **Bo Abrahamsen:** Writing – review and editing; resources; supervision. **Thomas Fuerst:** Data curation; writing – review and editing. **Marc Debois:** Conceptualization; funding acquisition; writing – review and editing. **Dirk Vandermeulen:** Supervision; writing – review and editing. **Cesar Libanati:** Conceptualization; data curation; funding acquisition; supervision; writing – review and editing.

Disclosures

Joeri Nicolaes, Marc Debois, Cesar Libanati: Employee and stock ownership, UCB Pharma. Joeri Nicolaes is involved in a patent (WO2019/106061). Michael Kriegbaum Skjødt: Institutional research grants from UCB/Amgen and Region Zealand Health Scientific Research Foundation (funds paid to the institution); support from the University of Southern Denmark (PhD scholarship), and UCB (educational grant and personal speakers fee) outside the submitted work; board member of the Danish Bone Society, and member of working groups in the Danish Bone Society and the European Calcified Tissue Society. Bo Abrahamsen: Speakers fees/consulting fees from UCB, MSD, Amgen, Kyowa-Kirin and Pharmacosmos. Institutional research grants from Novartis, UCB, Kyowa-Kirin and Pharmacosmos. Thomas Fuerst: Employee and stock ownership, Clario. Steven Raeymaeckers, Christopher Dyer Smith, and Dirk Vandermeulen have nothing to disclose. Michael Kriegbaum Skjødt and Christopher Dyer Smith had full access to individual-level data of all subjects in the validation set. Under Danish law sharing of these individual-level data is not possible. Other data generated or analyzed during the study are available from the corresponding author by request.

Peer Review

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1002/jbmr.4916.

15234681, 0, Downloaded from https://asbmr.onlinelibrary.wiley.com/doi/10.1002/jbmr.4916, Wiley Online Library on [17/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Data Availability Statement

Michael Kriegbaum Skjødt and Christopher Dyer Smith had full access to individual-level data of all subjects in the validation set. Under Danish law sharing of these individual-level data is not possible. Other data generated or analyzed during the study are available from the corresponding author by request.

References

1. Reginster JY, Burlet N. Osteoporosis: a still increasing prevalence. Bone. 2006;38(2):4–9.
2. Ström O, Borgström F, Kanis JA, et al. Osteoporosis: burden, health care provision and opportunities in the EU. Arch Osteoporos. 2011; 6(1):59–155.
3. Compston JE, McClung MR, Leslie WD. Osteoporosis. Lancet. 2019; 393:364–376.
4. Cauley JA, Thompson DE, Ensrud KC, Scott JC, Black D. Risk of mortality following clinical fractures. Osteoporos Int. 2000;11(7):556–561.
5. Chotiyarnwong P, McCloskey EV, Harvey NC, et al. Is it time to consider population screening for fracture risk in postmenopausal women? A position paper from the international osteoporosis foundation epidemiology/quality of life working group. Arch Osteoporos. 2022;17(1):1–24.
6. Cooper C, Atkinson EJ, MichaelO'Fallon W, Melton JL III. Incidence of clinically diagnosed vertebral fractures: a population-based study in Rochester, Minnesota, 1985-1989. J Bone Miner Res. 1992;7(2): 221–227.
7. Bhargavan M, Kaye AH, Forman HP, Sunshine JH. Workload of radiologists in United States in 2006–2007 and trends since 1991–1992. Radiology. 2009;252(2):458–467.
8. Bruls RJ, Kwee RM. Workload for radiologists during on-call hours: dramatic increase in the past 15 years. Insights Imaging. 2020;11(1): 1–7.
9. McDonald RJ, Schwartz KM, Eckel LJ, et al. The effects of changes in utilization and technological advancements of cross-sectional imaging on radiologist workload. Acad Radiol. 2015;22(9):1191–1198.
10. Bartalena T, Rinaldi MF, Modolon C, et al. Incidental vertebral compression fractures in imaging studies: lessons not learned by radiologists. World J Radiol. 2010;2(10):399e404.
11. Mitchell RM, Jewell P, Javaid MK, McKean D, Ostlere SJ. Reporting of vertebral fragility fractures: can radiologists help reduce the number of hip fractures? Arch Osteoporos. 2017;12(1):1–6.
12. Link TM. Osteoporosis imaging: state of the art and advanced imaging. Radiology. 2012;263(1):3.
13. Genant HK, Wu CY, Van Kuijk C, Nevitt MC. Vertebral fracture assessment using a semiquantitative technique. J Bone Miner Res. 1993; 8(9):1137–1148.
14. Ferrar L, Jiang G, Adams J, Eastell R. Identification of vertebral fractures: an update. Osteoporos Int. 2005;16(7):717–728.
15. Buckens CF, de Jong PA, Mol C, et al. Intra and interobserver reliability and agreement of semiquantitative vertebral fracture assessment on chest computed tomography. PloS One. 2013;8(8):e71204.
16. Ferrar L, Jiang G, Schousboe JT, DeBold CR, Eastell R. Algorithm-based qualitative and semiquantitative identification of prevalent vertebral fracture: agreement between different readers, imaging modalities, and diagnostic approaches. J Bone Miner Res. 2008;23(3):417–424.
17. Lentle B, Koromani F, Brown JP, et al. The radiology of osteoporotic vertebral fractures revisited. J Bone Miner Res. 2019;34(3):409–418.
18. Aggarwal V, Maslen C, Abel RL, et al. Opportunistic diagnosis of osteoporosis, fragile bone strength and vertebral fractures from routine CT scans; a review of approved technology systems and pathways to implementation. Ther Adv Musculoskelet Dis. 2021;13: 1759720X211024029.
19. Smets J, Shevroja E, Hügle T, Leslie WD, Hans D. Machine learning solutions for osteoporosis—a review. J Bone Miner Res. 2021;36(5): 833–851.
20. Yilmaz EB, Buerger C, Fricke T, et al. Automated deep learning-based detection of osteoporotic fractures in CT images. International Workshop on Machine Learning in Medical Imaging. Cham: Springer; 2021 pp 376–385.
21. Husseini M, Sekuboyina A, Loeffler M, Navarro F, Menze BH, Kirschke JS. Grading loss: A fracture grade-based metric loss for vertebral fracture detection. In Martel AL et al., eds. Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. MICCAI 2020. Lecture Notes in Computer Science, vol. 12266. Lima, Peru: Springer; 2022 pp 733–742.
22. Valentinitsch A, Trebeschi S, Kaesmacher J, et al. Opportunistic osteoporosis screening in multi-detector CT images via local classification of textures. Osteoporos Int. 2019;30(6):1275–1285.
23. Tomita N, Cheung YY, Hassanpour S. Deep neural networks for automatic detection of osteoporotic vertebral fractures on CT scans. Comput Biol Med. 2018;1(98):8–15.
24. Burns JE, Yao J, Summers RM. Vertebral body compression fractures and bone density: automated detection and classification on CT images. Radiology. 2017;284(3):788.
25. Baum T, Bauer JS, Klinder T, et al. Automatic detection of osteoporotic vertebral fractures in routine thoracic and abdominal MDCT. Eur Radiol. 2014;24(4):872–880.
26. Roux C, Rozes A, Reizine D, et al. Fully automated opportunistic screening of vertebral fractures and osteoporosis on more than 150 000 routine computed tomography scans. Rheumatology. 2022;61(8):3269–3278.
27. Kolanu N, Silverstone EJ, Ho BH, et al. Clinical utility of computer-aided diagnosis of vertebral fractures from computed tomography images. J Bone Miner Res. 2020;35(2312):2307–2312.
28. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553): 436–444.
29. Zhou SK, Greenspan H, Davatzikos C, et al. A review of deep learning in medical imaging: imaging traits, technology trends, case studies with progress highlights, and future promises. Proc IEEE Inst Electr Electron Eng. 2021;109(5):820–838.
30. Flahault A, Cadilhac M, Thomas G. Sample size calculation should be performed for design accuracy in diagnostic test studies. J Clin Epidemiol. 2005;58(8):859–862.
31. Bossuyt PM, Reitsma JB, Bruns DE, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. Clin Chem. 2015;61(12):1446–1452.
32. Koenig M, Spindler W, Rexilius J, Jomier J, Link F, Peitgen HO. Embedding VTK and ITK into a visual programming and rapid prototyping platform. Medical Imaging 2006: Visualization, Image-Guided Procedures, and Display, vol. 6141. San Diego, CA: SPIE; 2006 pp 796–806.
33. Skjødt MK, Nicolaes J, Smith CD, et al. Fracture risk in men and women with vertebral fractures identified opportunistically on routine CT scans and not treated for osteoporosis: an observational cohort study. JBMR Plus. 2023;7:e10736.
34. Nicolaes J, Raeymaeckers S, Robben D, et al. Detection of vertebral fractures in CT using 3D convolutional neural networks. In Cai Y, Wang L, Audette M, Zheng G, Li S, eds. Computational Methods and Clinical Applications for Spine Imaging. CSI 2019. Lecture Notes in Computer Science, vol. 11963. Shenzhen, China: Springer; 2020 pp 3–14.
35. Payer C, Štern D, Bischof H, Urschler M. Integrating spatial configuration into heatmap regression based CNNs for landmark localization. Med Image Anal. 2019;1(54):207–219.
36. Sekuboyina A, Husseini ME, Bayat A, et al. VerSe: a vertebrae labelling and segmentation benchmark for multi-detector CT images. Med Image Anal. 2021;1(73):102166.
37. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. J Mach Learn Res. 2011;12:2825–2830.
38. Abadi M, Agarwal A, Barham P, et al. TensorFlow: large-scale machine learning on heterogeneous systems; 2015. Software available from tensorflow.org.
39. Lowekamp BC, Chen DT, Ibáñez L, Blezek D. The design of SimpleITK. Front Neuroinform. 2013;30(7):45.
40. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb). 2012;22(3):276–282.

15234681, 0, Downloaded from https://asbmr.onlinelibrary.wiley.com/doi/10.1002/jbmr.4916, Wiley Online Library on [17/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License