Osteoporosis International (2024) 35:143–152
https://doi.org/10.1007/s00198-023-06903-7

ORIGINAL ARTICLE

External validation of a convolutional neural network algorithm for opportunistically detecting vertebral fractures in routine CT scans

Joeri Nicolaes[1,2] · Yandong Liu[3] · Yue Zhao[4] · Pengju Huang[5] · Ling Wang[3] · Aihong Yu[5] · Jochen Dunkel[6] · Cesar Libanati[2] · Xiaoguang Cheng[3]

Received: 20 February 2023 / Accepted: 29 August 2023 / Published online: 7 September 2023

Abstract

Summary The Convolutional Neural Network algorithm achieved a sensitivity of 94% and specificity of 93% in identifying scans with vertebral fractures (VFs). The external validation results suggest that the algorithm provides an opportunity to aid radiologists with the early identification of VFs in routine CT scans of abdomen and chest.

Purpose To evaluate the performance of a previously trained Convolutional Neural Network (CNN) model to automatically detect vertebral fractures (VFs) in CT scans in an external validation cohort.

Methods Two Chinese studies and clinical data were used to retrospectively select CT scans of the chest, abdomen and thoracolumbar spine in men and women aged ≥50 years. The CT scans were assessed using the semiquantitative (SQ) Genant classification for prevalent VFs in a process blinded to clinical information. The performance of the CNN model was evaluated against reference standard readings by the area under the receiver operating characteristics curve (AUROC), accuracy, Cohen's kappa, sensitivity, and specificity.

Results A total of 4,810 subjects were included, with a median age of 62 years (IQR 56-67), of which 2,654 (55.2%) were females. The scans were acquired between January 2013 and January 2019 on 16 different CT scanners from three different manufacturers. 2,773 (57.7%) were abdominal CTs. A total of 628 scans (13.1%) had ≥1 VF (grade 2-3), representing 899 fractured vertebrae out of a total of 48,584 (1.9%) visualized vertebral bodies. The CNN's performance in identifying scans with ≥1 moderate or severe fractures achieved an AUROC of 0.94 (95% CI: 0.93-0.95), accuracy of 93% (95% CI: 93%-94%), kappa of 0.75 (95% CI: 0.72-0.77), a sensitivity of 94% (95% CI: 92-96%) and a specificity of 93% (95% CI: 93-94%).

Conclusion The algorithm demonstrated excellent performance in the identification of vertebral fractures in a cohort of chest and abdominal CT scans of Chinese patients ≥50 years.

Keywords Convolutional neural network · Osteoporosis · Predictive value of tests · Tomography, X-Ray computed · Vertebral body

Introduction

Osteoporosis affects approximately 200 million people globally, resulting in more than 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. It is estimated that only one out of three VFs comes to clinical attention [3] and underreporting of VFs is a worldwide problem [4]. Radiologists apply different protocols for reading VFs that can be categorized as qualitative, quantitative, or semiquantitative (SQ) assessments. Several studies have shown that inter- and intra-reader variability is significant for the various reading standards and across

Joeri Nicolaes and Yandong Liu contributed equally to this work.

✉ Joeri Nicolaes
joeri.nicolaes@kuleuven.be

[1] Department of Electrical Engineering (ESAT), Center for Processing Speech and Images, KU Leuven, Leuven, Belgium

[2] UCB Pharma, Brussels, Belgium

[3] Department of Radiology, Beijing Jishuitan Hospital, Beijing 100035, China

[4] Department of Radiology, Qingdao Fuwaicardiovascular Hospital, Qingdao 26600, China

[5] Department of Radiology, Beijing Anding Hospital, Beijing 100120, China

[6] UCB Pharma, Monheim, Rhein, Germany

modalities [5, 6]. The growing number of abdominal and chest CT scans and associated radiologist workload [7] provides an opportunity to aid radiologists with the identification and reporting of VFs using artificial intelligence algorithms.

Over the last decades, Convolutional Neural Networks (CNN) have been successfully applied to detection and segmentation tasks in medical image analysis [8]. Existing VF detection methods vary in their degree of automation, supported modality, modeling approach, and maturity. While most methods are fully automated, none of them independently diagnose VFs without confirmation from a clinician and hence they operate as computer-aided support systems. The published machine learning methods for VF detection are predominantly applied to CT, DXA, and lateral radiographs, yet they all leverage information from 2D images only (i.e., sagittal reformations in the case of CT) [9]. Modeling approaches range from segmentation of vertebral bodies followed by height measurements to end-to-end methods automatically scoring an image as containing VFs or not [10–15]. Evidence of general applicability is limited for most of the above VF detection methods, which present results from cross-validation studies using relatively small sample sizes acquired at a single center (order of magnitude of 100 samples). Some methods were evaluated further with retrospective diagnostic validation studies on external data from multiple centers but again on data sets of a few hundred [14, 16, 17] to almost 1,700 samples [18]. Finally, one study discussed the challenges and opportunities of integrating two devices that were approved for clinical use for opportunistically screening CT scans into the osteoporosis care pathway in the context of the UK National Health Service [19].

The purpose of our study was therefore to evaluate the performance of a previously trained Convolutional Neural Network (CNN) model to automatically detect VFs in CT scans in an external validation cohort and to investigate its potential for helping radiologists to identify VFs on routine CT scans.

Materials and Methods

Study design and cohort

We retrospectively collected 5,195 CT scans: 2,419 abdominal CT images and 2,036 chest CT images originating from prior community health screening studies, and 740 thoracic/lumbar spine CT images from Beijing Jishuitan Hospital (Fig. 1). 2,419 abdominal CT scans were randomly sampled from the China Action on Spine and Hip (CASH) study, an epidemiology study that recruited 3,457 subjects across seven Chinese provinces and performed lumbar spine Quantitative Computed Tomography (QCT) between June 2013 and March 2017. In the CASH study, Bone Mineral Density (BMD) measurements were computed on the QCT scans and scout views were reviewed for prevalent VFs to determine the prevalence of osteoporosis and evaluate the association between VFs and BMD [20]. 2,036 chest CT images in which vertebrae

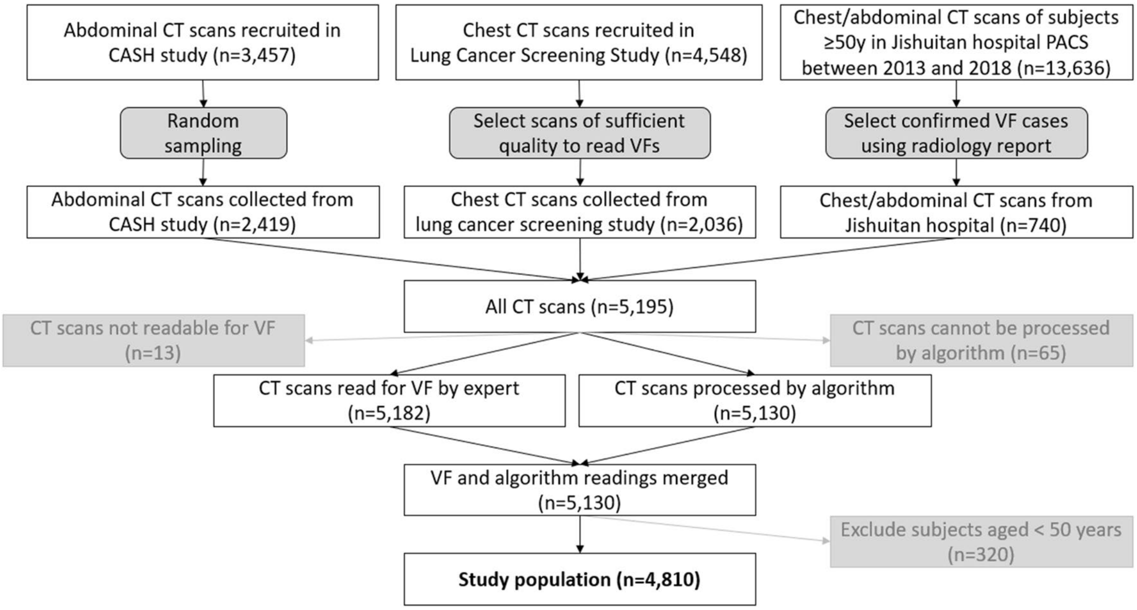
Abdominal CT scans recruited in CASH study (n=3,457)
Chest CT scans recruited in Lung Cancer Screening Study (n=4,548)
Chest/abdominal CT scans of subjects ≥50y in Jishuitan hospital PACS between 2013 and 2018 (n=13,636)
Random sampling
Select scans of sufficient quality to read VFs
Select confirmed VF cases using radiology report
Abdominal CT scans collected from CASH study (n=2,419)
Chest CT scans collected from lung cancer screening study (n=2,036)
Chest/abdominal CT scans from Jishuitan hospital (n=740)
All CT scans (n=5,195)
CT scans not readable for VF (n=13)
CT scans cannot be processed by algorithm (n=65)
CT scans read for VF by expert (n=5,182)
CT scans processed by algorithm (n=5,130)
VF and algorithm readings merged (n=5,130)
Exclude subjects aged < 50 years (n=320)
Study population (n=4,810)

Fig. 1 Flowchart of the study. VF = vertebral fracture. CASH = China action on spine and hip. PACS = picture archive and communications system

could be identified by naked eye were selected, excluding 2,512 low-dose CT exams due to low-resolution, from the Beijing Lung Cancer Early Screening and Early Treatment program, an epidemiological study that recruited 4,548 subjects in the Beijing community and performed chest CT imaging between August 2013 and March 2014. In the original study, lateral chest CT images were reviewed for lumbar spondylolisthesis (LS) of L1 to L5 to assess the prevalence of LS [21]. The 740 thoracic/lumbar spine CT scans were collected from Beijing Jishuitan Hospital by oversampling for fracture cases to ensure a validation set with sufficient VFs. LY queried the hospital Picture Archiving and Communication System (PACS) for CT exams performed between January 2013 and December 2018 and searched within the radiology reports using the keywords "fracture" and "wedge deformity". Cases belonging to women and men aged 50 years or older at the time of their CT scan with osteoporotic vertebral fracture or wedge deformity were included, while cases of subjects <50 years old, with traumatic fracture, pathological fracture and metal internal fixation of spine were excluded. CT scans with a maximum slice thickness of 2 mm that had a quality deemed acceptable to read VFs by a radiologist were included. All CT scans were anonymized before processing. We estimated that a sample size of 624 subjects with VFs was sufficient to measure a sensitivity of 80% or more, which we were expecting to meet in our study population assuming a VF prevalence of 15% [22].

VF detection model development

This study evaluated a previously developed CNN algorithm that automatically processes a CT scan blinded to clinical information and outputs a list of vertebrae with a VF grade associated with every level identified in the scan (Fig. 2). The algorithm is composed of two models: a model that estimates the SQ grade (i.e., SQ0=normal, SQ1=mild, SQ2=moderate, and SQ3=severe VF) for every vertebra visible in the CT scan, and a model that identifies its anatomical level (i.e., T1, ..., L5). The VF detection model was previously trained by JN on a private dataset of 666 CT scans from the Universitair Ziekenhuis (UZ) Brussel. The training set comprised of abdominal and chest CT scans of subjects aged 50 years or older at the time of the scan with a maximum slice thickness of 3mm. VF readings were defined in the training set following the Genant SQ method involving an external radiology service (Clario, USA). The training set was balanced for the presence of VFs at subject-level (VF prevalence of 55% at subject-level), oversampling VFs to ensure that enough VFs were present at every vertebral level for every SQ grade (VF prevalence of 12% at vertebral-level). The VF detection model consists of a 3D CNN model that outputs SQ grade scores for every voxel in the CT image and a post-processing step to aggregate the voxel-level scores to vertebral- and subject-level outcomes [23]. We applied an ensemble of three models, trained on different subsets of the UZ Brussel dataset, by averaging the vertebral-level scores for each SQ grade across all models.

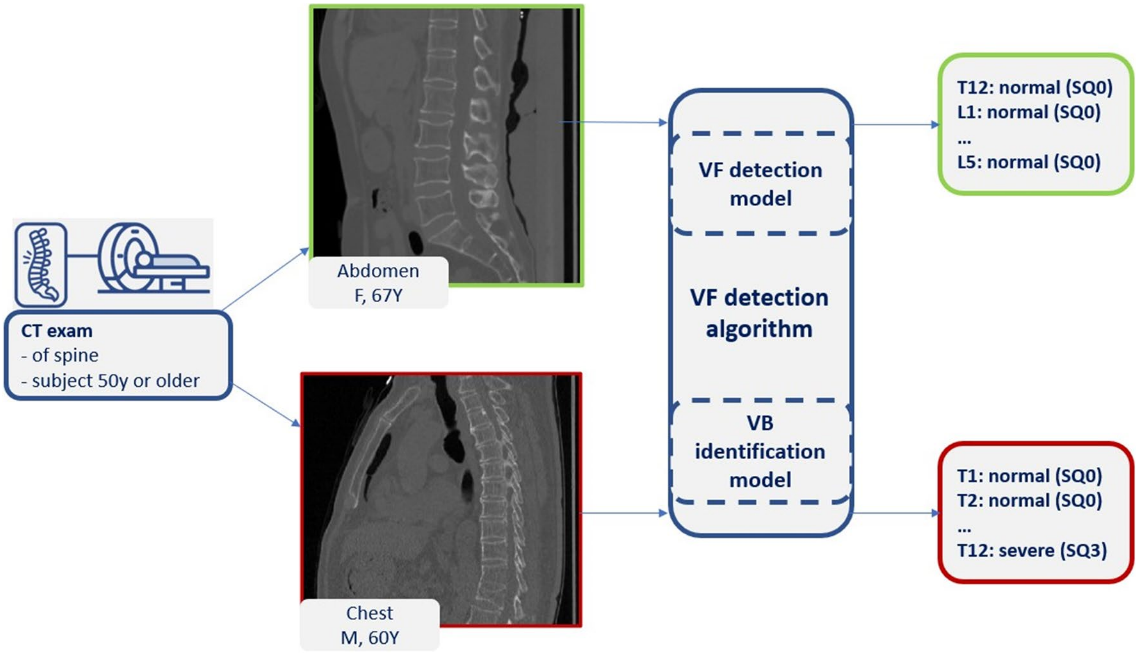

Fig. 2 Illustration of the machine learning algorithm processing abdominal/chest CT scans. CT = computed tomography. VF = vertebral fracture. VB = vertebral body. SQ = semiquantitative grade (Genant). Vertebra levels as T1 = first thoracic vertebra, …, L5 = fifth lumbar vertebra

This approach is similar to asking three experts to independently grade each vertebra in a CT scan and average their results. The vertebra identification model was previously trained by Payer et al. on the Verse dataset [24, 25]. The VF detection algorithm processes any spine CT exam, irrespective of the exam type, number of visible vertebrae, number of slices and acquisition characteristics (e.g., with or without contrast, different convolution kernels, etc.). The CT images underwent the following preprocessing steps: resample to 1x1x1mm spacing, clip CT intensities below -1,000 and above 2,000 Hounsfield units and normalize intensities to zero mean and unit standard deviation. The algorithm was executed on a server using one NVidia GTX 1080 Ti GPU card and leveraged the Tensorflow [26] and SimpleITK [27] Python software packages.

Reference standard reading

All CT scans were read twice by radiologists in a process blinded to clinical information. First, one sub-specialist radiologist (XC) with more than 20 years' experience identified and graded fractured vertebrae from the lateral CT scout view according to Genant's SQ method [28]. The SQ method is recommended by most societies such as ISCD, IOF and the European Society of Skeletal Radiology [29] and commonly applied in research studies as a gold standard. Second, the study cohort was randomly split into three subsets that were each assigned to one reader. Three sub-specialist radiologists with less than five years' experience (YL, PH, YZ) read individual vertebrae applying the SQ method on sagittal slices of the CT scans, blinded to the first readings. In the event of a disagreement between the first and second readings, the final grade was determined by a consensus review involving all four readers.

Statistical analyses

The performance of the algorithm was evaluated against reference standard readings for primary outcomes SQ23 (grade 0-1 vs. grade 2-3) and secondary outcomes SQ123 (normal vs. grade 1-3) at subject- and vertebral-level by the area under the Receiver Operating Characteristics curve (AUROC), accuracy, Cohen's kappa, sensitivity, specificity, and positive and negative predictive values (PPV and NPV). Accelerated bootstrapping with bias-correction using 1,000 repetitions was used to construct the 95% confidence intervals (CI) and box plots. A threshold of 0.5 was used for all cut-off metrics. Groups were compared by two-tailed Student t-tests for continuous and χ^2-tests for categorical data using significance level alpha=.05. Subject- and vertebral-level analysis was performed when both the human and model readings were available. Statistical analyses were performed using Scikit-learn v1.0 [30].

Results

Study data

Of the 5,195 CT scans collected in this study, the readers excluded 13 CT scans due to poor image quality. The same 13 CT scans and 52 other CT scans could not be read by the algorithm due to missing slices in the DICOM CT series. Both the reference standard and algorithm readings were available for a total of 5,130 CT scans belonging to unique individuals. We excluded 320 CT scans of subjects aged <50 years (Fig. 1). As a result, 4,810 CT scans were eligible for analysis. The study population had a median age of 62 years (IQR: 56-67), of which 2,654 (55.2%) were females. The baseline characteristics of our study population is shown for the 'No VF (normal or mild VF)' and 'VF (moderate or severe VF)' sub-groups in Table 1. As expected, the prevalence of VFs increased with age, with a median age of 61 and 69 years in the 'No VF' and 'VF' groups respectively (p-value <.001, Table 1a). The sex and number of vertebrae visible in the CT scan differed significantly between both groups (both p-values <.001, Table 1a). The median number of vertebrae visible was 13 (IQR: 12-13) and 9 (IQR: 8-10) in chest and abdominal CT scans respectively. 2,773 (57.7%) were abdominal CTs. The 'VF' group contained a higher proportion of abdominal exams than the 'No VF' group (89% vs. 53% respectively, p-value <.001, Table 1a). While the median age of female and male subjects was similar (62 and 61 years respectively), there were significantly more older women (specifically in age group [70,79]). The abdominal CT scans belonged to older subjects (median age 65 vs. 58 years in chest exams, p-value <.001, Table 1b), with major differences in age groups [50,59] and [70,89] (Table 1b). The CT scans were acquired at eight different institutions on 16 CT scanners from three different manufacturers. The peak kilo voltage was the same for all scans (120 kVp) and the x-ray tube current ranged from 50 to 250 mAs. The slice thickness was 0.625mm (13%), 1mm (34%) and 1.25mm (53%). Six convolution kernels were used (Soft, Standard, B30F, Lung, FC03 and Bone).

The VF reference standard readings resulted in a total of 628 scans (13.1%) with ≥1 moderate or severe VFs (grades 2-3), representing 899 fractured vertebrae out of a total of 48,584 (1.9%) visualized vertebral bodies. A total of 1,622 scans (33.7%) with ≥1 VF (grades 1-3) were found, representing 2,623 fractured vertebrae out of a total of 48,584 (5.4%) visualized vertebral bodies. The cumulative fracture grade (i.e., the sum of all VF grades in a scan) had a median of 2 (IQR: 1-3) in the subset of positive cases. Almost two-thirds of these scans contained only mild VFs (994 out of 1622), of which 687 scans contained

Table 1 Baseline characteristics of study participants

(a) Participants stratified in 'no VF' (SQ 0-1) and VF (SQ 2-3) groups as per reference standard readings: both groups show significant differences in sex, age (as expected) and CT scan characteristics. Median data are provided with IQR in parentheses.

Characteristic	No VF (SQ 0-1) (*n*=4,182)	VF (SQ 2-3) (*n*=628)	*P* value
Subject			
Women; N (%)	2,197 (52.5%)	457 (72.8%)	<.001
Median age, y (IQR)	61 (56-66)	69 (63-76)	<.001
CT scan			
Median number of visible vertebrae (IQR)	11 (9-12)	8 (7-10)	<.001
Abdominal exams; N (%)	2,214 (52.9%)	559 (89.0%)	<.001
Scans performed on GE; N (%)	3,094 (74.0%)	97 (15.5%)	<.001
Scans performed on Siemens; N (%)	884 (21.1%)	68 (10.8%)	…
Scans performed on Toshiba; N (%)	204 (4.9%)	463 (73.7%)	…

(b) Participants stratified by age groups [50-59], [60-69], [70-79], [80-89]: the study population contains significantly more women aged ≥70y and significantly more abdomen exams of subjects aged ≥70y. Subjects of 90 years or older are not shown in this table (N<5).

Characteristic	Age [50-59y]	Age [60-69y]	Age [70-79y]	Age [80-89y]	Median age, y (IQR)	P value
Women; N (%)	976 (37%)	1,130 (43%)	464 (18%)	81 (3%)	62 (57-68)	<.001
Men; N (%)	900 (42%)	942 (44%)	267 (12%)	46 (2%)	61 (56-66)	…
Abdominal exams; N (%)	675 (24%)	1,260 (45%)	714 (26%)	120 (4%)	65 (60-71)	<.001
Chest exams; N (%)	1,201 (59%)	812 (40%)	17 (1%)	7 (0%)	58 (55-62)	…

IQR = interquartile range. GE = General Electric. VF = vertebral fracture. SQ = semiquantitative grade

only a single mild VF. 41% (666 out of 1,622) and 75% (472 out of 628) of the SQ123 and SQ23 cases, respectively, originated from Jishuitan hospital (Fig. 1). The readers additionally annotated the presence of Schmorl's nodes, which were found in 17% (n=819) of all CT scans. Figure 3 shows the number of VFs for every SQ grade, the proportion of fractured vertebrae, and the total number of visible vertebrae from T1 to L5 in our study population.

Diagnostic performance of VF detection model vs. reference standard readings

The metrics for the evaluation of the diagnostic performance of the VF detection model versus reference standard readings are shown in Table 2. The SQ23 subject-level performance in differentiating normal/mild from moderate/severe VFs shows an AUROC of 0.938 (95% CI: 0.928-0.947), Cohen's kappa of 0.749 (95% CI: 0.722-0.774), accuracy of 93.3% (4,490 of 4,810), sensitivity of 94.4% (593 of 628), specificity of 93.2% (3,897 of 4,182), PPV of 67.5% (593 of 878) and NPV of 99.1% (3,897 of 3,932). The SQ23 vertebral-level performance for identifying grade 2-3 VFs has

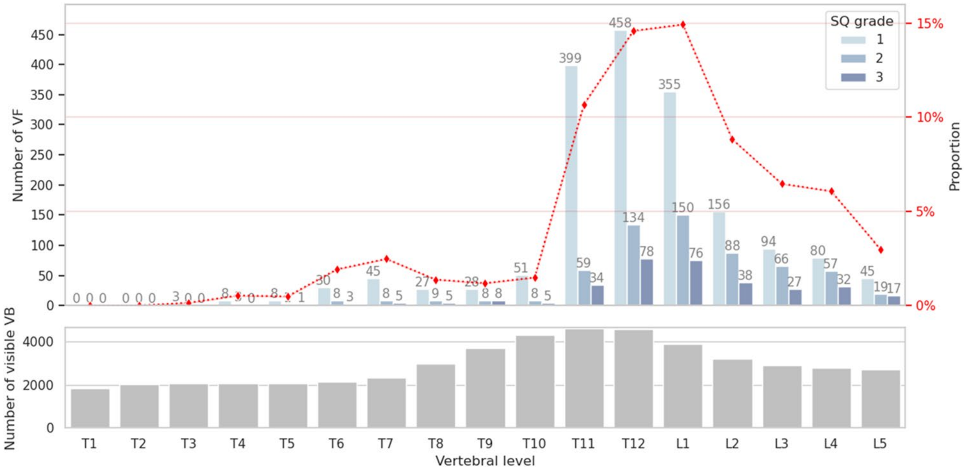

Fig. 3 Distribution of vertebrae and VF across vertebral levels T1 to L5 (on the x-axis). The top panel shows on the left y-axis, the count of VF stratified per SQ grade with a bar plot (SQ1 or mild VF in light blue, SQ2 or moderate VF in blue and SQ3 or severe VF in dark blue), and on the right y-axis the proportion of VF with a line plot (in red). The bottom panel shows the total number of visible vertebrae. VF = vertebral fracture. VB = vertebral body. SQ = semiquantitative grade

Table 2 Diagnostic performance of VF detection model vs. reference standard readings. The metrics are stratified by outcome SQ23 (normal and mild vs. grade 2-3) and SQ123 (normal vs. grade 1-3), and level (subject and vertebral). The depicted numbers are point estimates with 95% CI between parentheses, all CI's have been generated using a bias-corrected and accelerated bootstrapping method (1,000 iterations)

(a) SQ23 (normal/mild vs. moderate/severe)		
Metric	Subject-level	Vertebral-level
AUROC	0.938 (0.928-0.947)	0.932 (0.921-0.943)
Accuracy	0.933 (0.926-0.940)	0.988 (0.987-0.989)
Kappa	0.749 (0.722-0.774)	0.728 (0.706-0.746)
Sensitivity	0.944 (0.923-0.960)	0.874 (0.853-0.896)
Specificity	0.932 (0.925-0.939)	0.990 (0.989-0.991)
PPV	0.675 (0.646-0.708)	0.632 (0.605-0.659)
NPV	0.991 (0.988-0.994)	0.998 (0.997-0.998)
(b) SQ123 (normal vs. mild/moderate/severe) outcome		
Metric	Subject-level	Vertebral-level
AUROC	0.781 (0.768-0.794)	0.783 (0.773-0.792)
Accuracy	0.831 (0.821-0.841)	0.967 (0.966-0.969)
Kappa	0.598 (0.574-0.621)	0.639 (0.623-0.656)
Sensitivity	0.626 (0.602-0.648)	0.576 (0.557-0.595)
Specificity	0.935 (0.926-0.943)	0.990 (0.989-0.991)
PPV	0.831 (0.809-0.851)	0.762 (0.744-0.782)
NPV	0.831 (0.819-0.843)	0.976 (0.975-0.977)

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

an AUROC of 0.932 (95% CI: 0.921-0.943), Cohen's kappa of 0.728 (95% CI: 0.706-0.746), accuracy of 98.8% (48,013 of 48,584), sensitivity of 87.4% (786 of 899), specificity of 99.0% (47,227 of 47,685), PPV of 63.2% (786 of 1,244) and NPV of 99.8% (47,227 of 47,340). In a sensitivity analysis, we found that one third of the subject-level false positives (87 of 285) were explained by errors involving the first and last visible vertebrae in the CT scan. This finding is consistent with the methodology of the VF detection algorithm: every vertebra is analyzed by considering its superior and inferior neighboring vertebrae (mimicking the information used by a human reader), and an absent top or bottom vertebra increases the ambiguity of the reading. Additionally, we found that one out of five false positive (56 of 285) scans contained Schmorl's nodes. The SQ123 subject-level performance in differentiating normal from grade 1-3 VFs shows an AUROC of 0.781 (95% CI: 0.768-0.794), Cohen's kappa of 0.598 (95% CI: 0.574-0.621), accuracy of 83.1% (3,997 of 4,810), sensitivity of 62.6% (1,016 of 1,622), specificity of 93.5% (2,981 of 3,188), PPV of 83.1% (1,016 of 1,223) and NPV of 83.1% (2,981 of 3,587). The SQ123 vertebral-level performance for grade 1-3 VFs show an AUROC of 0.783 (95% CI: 0.773-0.792), Cohen's kappa of 0.639 (95% CI: 0.623-0.656), accuracy of 96.7% (47,001 of 48,584), sensitivity of 57.6% (1,511 of 2,623), specificity of 99.0% (45,490 of 45,961), PPV of 76.2% (1,511 of 1,982) and NPV of 97.6% (45,490 of 46,602). The majority (586 of 606) of false negatives were mild VF cases; thus, the lower sensitivity can be attributed to missing mild VFs. For false positives, 31% (64 of 207) of the scans were attributable to mild VFs only. In a sensitivity analysis, we found that half of the subject-level false positives (103 of 207) and roughly 15% (90 of 606) of the subject-level false negatives were explained by errors involving the first and last visible vertebrae in the CT scan. Almost one third of the subject-level false positives (59 of 207) and 13% (79 of 606) of the subject-level false negatives occurred in scans with Schmorl's nodes present.

Figure 4 qualitatively illustrates the primary results for four CT scans that were selected from the study population by XC from the abdomen and chest sub-groups without knowledge of the algorithm outputs. The presented images contain one sagittal slice extracted from the CT scan and the algorithm heatmaps overlaid in color using ITK-SNAP [31]. Note that these slices are only used for visualization purposes; the reference standard readings and the algorithm outputs were generated using all 3D information contained within the CT scan. Figures (b) and (c) illustrate that all normal vertebrae present were correctly detected on the normal scans with medium (green) to high (red) confidence scores, except for the T1 vertebra which got detected with lower (blue) confidence in the chest scan (Figure c). Figures (d) and (e) illustrate that the algorithm confidently detected one severe VF and three moderate or severe VFs respectively with good (yellow) to high (red) confidence scores. The moderate L2 was missed (identified as a mild VF by the algorithm) in the chest scan with blue heatmap colors representing low algorithm scores for SQ2 and SQ3 (Fig. 4e). The lack of heatmap colors for the other vertebrae in sub-figures (d) and (e) demonstrate that the algorithm confidently identified those vertebrae as normal or mild VF, since the algorithm's confidence scores for a moderate or severe VF were below 0.05.

The machine learning algorithm required on average two minutes run-time per scan using GPU acceleration.

Discussion

We retrospectively collected CT scans of 4,810 subjects from eight institutions across China. Expert radiologists annotated 48,584 vertebral bodies in this study population using the Genant SQ assessment method. The Convolutional Neural Network algorithm agreed substantially with the expert readers, reaching a Cohen's kappa of 0.75, similar to the agreement between readers reported by Buckens et al. [5]. The algorithm's sensitivity of 94%

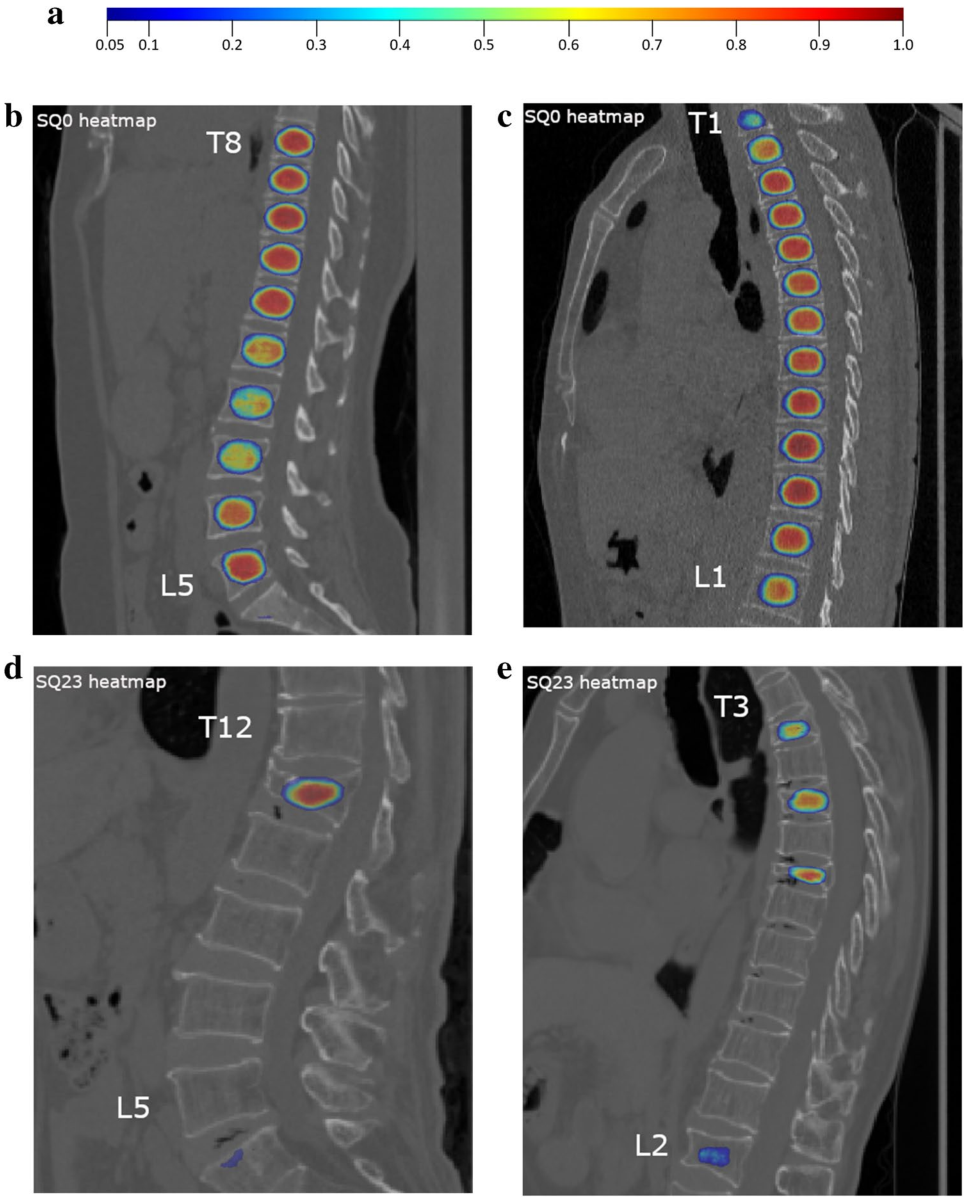

Fig. 4 Qualitative analysis of the primary results: visualization of algorithm heatmaps for four CT scans from our study population. The presented images contain one sagittal slice extracted from the CT scan and the algorithm heatmaps in ITK-SNAP [31]. CT image intensities are clipped below -1,000 and above 2,000 Hounsfield units and the heatmaps show the algorithm's confidence scores between 0.05 and 1.0. The heatmaps illustrate that the algorithm identifies the location of the vertebral bodies with heatmap colors representing the algorithm's confidence for the presence of a SQ grade shown in figure (**a**). Abdominal (left) and chest (right) CT scans are displayed. The SQ0 (normal/no VF) heatmap is overlaid in figures (**b**) and (**c**), hence the heatmap colors represent the algorithm's confidence for the presence of a normal vertebra. The SQ2 and SQ3 (moderate or severe VF) heatmaps are overlaid in figures (**d**) and (**e**), hence the heatmap colors represent the algorithm's confidence for the presence of a moderate or severe VF. The sex, age, reference standard reading and vertebral-level agreement for row three are depicted under each sub-figure. Note that these slices are only used for visualization purposes; the reference standard readings and the algorithm outputs were generated using all 3D information contained within the CT scan. The authors welcome the interested reader to get in contact to review these CT scans and algorithm outputs in full detail. F = female. M = male. SQ = semiquantitative grade. Y = years. TP = true positive. FN = false negative. (**a**) Heatmap colors representing the algorithm's confidence scores from 0.05 (blue) to 1.0 (red). (**b**) F, 53Y, normal. (**c**) M, 57Y, normal. (**d**) M, 78Y, severe L1 (TP). (**e**) F, 71Y, moderate T4 (TP), moderate T6 (TP), severe T8 (TP), moderate L2 (FN*). *The algorithm's confidence scores were highest for SQ1 (mild VF)

for identifying moderate and severe VFs suggests that the application of the CNN algorithm would improve current clinical practice, where more than half of the prevalent VFs are missed on imaging exams [4]. When applied in a clinical alerting workflow to flag cases with moderate or severe fractures to the radiologist for follow-up, the algorithm would reduce the workload by 80% (878 of 4,180 scans were detected as positive by the algorithm) compared to a fully manual review of every spine-containing CT scan by a radiologist. The Positive Predictive Value (PPV) of 68% and Negative Predictive Value (NPV) of 99% suggest that the algorithm is more reliable in detecting negative (normal or mild) than in detecting positive (moderate or severe) scans. While this operating point implies that radiologists need to re-classify 1 out of 3 scans flagged for follow-up as normal, the high sensitivity implies that less than 6% of scans with VFs were missed by the algorithm (35 of 628).

We found that the number of CT scans with grade 1-3 VFs is almost three times the number of scans with grade 2-3 VFs (1,622 vs. 628). The performance of the model in identifying scans with one or more grade 1-3 VFs is inferior to its performance identifying grade 2-3 VFs in AUROC (78% vs. 94%), Cohen's kappa (0.60 vs. 0.75), and accuracy (83% vs. 93%). More specifically, the model identified mild VFs more conservatively with a lower sensitivity (63% vs. 94%) and a higher PPV (83% vs. 68%) when comparing the SQ123 results with the SQ23 results (Table 2). The vast majority (97%) of false negative scans involve only mild VFs. The drop in performance when including mild grade VFs has been reported by other studies [17]. The lower performance on mild VFs was expected as the algorithm was trained on and tested against mild VF readings that exhibit higher reader variability [5]. Importantly, previous studies have shown that the association between mild VFs and low BMD is low [32], hence the algorithm performs best on the clinically most important VFs.

The vertebral-level SQ23 results are similar to the subject-level results on most metrics, except for a drop in sensitivity to 87% (Table 2). For the SQ123 outcome, the vertebral-level results differ on multiple metrics from the subject-level results. The vertebral- and subject-level results cannot be readily compared, given that the baseline characteristics of these data sets are different. We argue that the vertebral-level results are important to verify the algorithm outputs against the reference standard readings and to provide more detailed insights into the algorithm outputs. However, the subject-level results are the most important clinical outcomes because they influence treatment decisions. Vertebral-level results must be interpreted with caution as mismatches between the vertebral levels identified by the model and reader are common (perfect correspondences were found in 3088 or 64% of all scans). While such vertebral-level mismatches impact the vertebral-level results, they generally do not change the subject-level results (e.g., if the algorithm detects T12 as a moderate VF while the reference standard reading was a moderate VF at L1 and normal vertebra at T12, the vertebral-level results would be impacted but the subject-level results remain the same). We conclude that the algorithm performs best on the clinically most important outcomes, i.e., subject-level VFs.

We found that our study population had a higher VF prevalence for grades 1-3 and lower VF prevalence for grades 2-3 compared to another study that retrospectively identified VFs in abdominal and chest CT scans (34% vs. 24% and 13% vs. 18% respectively in ours vs. Kolanu et al. [18]). Ignoring the statistical differences between both validation studies, we found that the subject-level results were marginally higher in our study than the other study, which evaluated a different algorithm, for the SQ123 outcome (sensitivity of 54% vs. 63% in our study, PPV of 69% vs. 83% in our study, and specificity of 92% vs. 94% in our study), and the SQ23 outcome (sensitivity of 65% vs. 94% in our study, PPV of 65% vs. 68% in our study, and specificity of 92% vs. 93% in our study). We note that the study had a smaller sample size than ours (1,696 vs 4,810), did not report vertebral-level results and applied a specific adjudication procedure to determine the reference standard readings, which involved unblinding the algorithm readings to the study readers. Notably, the adjudication procedure involved only reviewing the reference standard readings in the subset of scans where the algorithm and the first reader disagreed. Another study has previously reported validation results for identifying VFs by measuring the diagnostic performance of a different algorithm compared to human readers on 500 CT scans randomly sampled from a study population [16]. The study did not identify the VFs at the individual vertebral level, nor did it report the presence of mild (grade 1) fractures. In this different cohort, the authors reported a similar subject-level sensitivity of 94% (albeit with a larger 95% CI: 89-98% vs. 92-96% in our study) but a lower specificity of 65% (95% CI: 60-70%) vs. 93% (95% CI: 93-94%) in our study.

In this study, we retrospectively examined CT scans from two prior epidemiological studies and one hospital PACS in a cohort of Chinese subjects, which constitutes a selection bias. Furthermore, our study population contained other selection biases that may not be present in a wider abdominal and chest CT population (i.e., specific CT acquisition settings, correlations between CT exam type and number of VFs in different age groups, and correlations between VFs and CT scan characteristics). Future work should aim at prospectively studying CT scans in routine clinical cohorts with bigger sample sizes to further demonstrate the algorithm's reliability and investigate further whether the algorithm performs similarly in specific sub-groups (e.g., male vs. female, thoracic vs. lumbar vertebrae, …). A sensitivity analysis of vertebral-level results (for different SQ grades and different levels) would require a manual consensus review to ensure perfect correspondence between the vertebral-level readings of the algorithm and the reader. This study evaluated the performance of a machine learning algorithm compared to reference standard readings defined by two readers with a consensus review in case of disagreement; future work should study the diagnostic performance of the algorithm and a dozen or more readers against a gold standard reading, using two study arms, i.e., one with and one without an algorithm as computer-aided support.

In summary, the CNN algorithm demonstrated excellent performance in the identification of vertebral fractures in an external validation cohort of abdominal and chest CT scans of Chinese patients ≥50 years. As life expectancy increases and the elderly population grows, the number of patients receiving CT examinations of the abdomen or chest for various indications is increasing. Applying the algorithm to flag the presence of vertebral fractures for radiologist review offers the potential to improve on the identification and reporting of VFs in patients aged 50 years or older without overloading radiologists.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00198-023-06903-7.

Acknowledgements The authors would like to thank Prof. Glen Blake for reviewing the manuscript on grammar, punctuation, and spelling.

Funding This study has received funding by the National Natural Science Foundation of China (grant no. 81971617), Beijing Hospitals Authority Youth Programme (QML20200402), the National Key R&D Program of China (grant no.2020YFC2004902), Beijing Hospitals Authority Clinical Medicine Development of Special Funding Support (ZYLX202107) and UCB Pharma/Amgen Inc. The manuscript was shared with UCB Pharma/Amgen Inc. for courtesy review prior to submission. The other funders were not involved in the study.

Data availability Data generated or analyzed during the study are available from the corresponding author by request.

Declarations

Conflicts of interest None.

Ethical approval This study was approved by the ethics board of Beijing Jishuitan Hospital (approval number 201903-24).

Informed consent Written informed consent was obtained from individual participants in the China Action on Spine and Hip (CASH) study and the Beijing Lung Cancer Early Screening and Early Treatment program. Informed consent can be waived for those retrospectively collected cases from Beijing Jishuitan Hospital PACS.

References

1. Reginster JY, Burlet N (2006) Osteoporosis: a still increasing prevalence. Bone 38(2 Suppl 1):S4–S9. https://doi.org/10.1016/j.bone.2005.11.024
2. Borgström F, Karlsson L, Ortsäter G et al (2020) Fragility fractures in Europe: burden, management and opportunities. Arch Osteoporos 15:1–21
3. Cooper C, Atkinson EJ, O'Fallon WM, Melton LJ 3rd (1992) Incidence of clinically diagnosed vertebral fractures: a population-based study in Rochester, Minnesota, 1985-1989. J Bone Miner Res 7(2):221–227
4. Bartalena T, Rinaldi MF, Modolon C et al (2010) Incidental vertebral compression fractures in imaging studies: lessons not learned by radiologists. World J Radiol 2(10):399e404
5. Buckens CF, de Jong PA, Mol C et al (2013) Intra and interobserver reliability and agreement of semiquantitative vertebral fracture assessment on chest computed tomography. PLoS One 8(8):e71204. https://doi.org/10.1371/journal.pone.0071204
6. 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
7. McDonald RJ, Schwartz KM, Eckel LJ, et al (2015) The effects of changes in utilization and technological advancements of cross-sectional imaging on radiologist workload. Acad Radiol 1;22(9):1191-8. https://doi.org/10.1016/j.acra.2015.05.007
8. Litjens G, Kooi T, Bejnordi BE et al (2017) A survey on deep learning in medical image analysis. Med Image Anal 42:60–88. https://doi.org/10.1016/j.media.2017.07.005
9. Smets J, Shevroja E, Hügle T, Leslie WD, Hans D (2021) Machine Learning Solutions for Osteoporosis-A Review. J Bone Miner Res 36(5):833–851. https://doi.org/10.1002/jbmr.4292
10. Yilmaz EB, Buerger C, Fricke T, et al (2021) Automated Deep Learning-Based Detection of Osteoporotic Fractures in CT Images. In: International Workshop on Machine Learning in Medical Imaging Springer. https://doi.org/10.1007/978-3-030-87589-3_39
11. Husseini M, Sekuboyina A, Loeffler M, Navarro F, Menze BH, Kirschke JS (2020) Grading loss: a fracture grade-based metric loss for vertebral fracture detection. In: International Conference on Medical Image Computing and Computer-Assisted Intervention Springer. https://doi.org/10.1007/978-3-030-59725-2_71
12. Valentinitsch A, Trebeschi S, Kaesmacher J et al (2019) Opportunistic osteoporosis screening in multi-detector CT images via local classification of textures. Osteoporos Int 30(6):1275–1285. https://doi.org/10.1007/s00198-019-04910-1
13. Tomita N, Cheung YY, Hassanpour S (2018) Deep neural networks for automatic detection of osteoporotic vertebral fractures on CT scans. Comput Biol Med 98:8–15. https://doi.org/10.1016/j.compbiomed.2018.05.011
14. Burns JE, Yao J, Summers RM (2017) Vertebral body compression fractures and bone density: automated detection and classification on CT images. Radiol 284(3):788. https://doi.org/10.1148/radiol.2017162100
15. Baum T, Bauer JS, Klinder T et al (2014) Automatic detection of osteoporotic vertebral fractures in routine thoracic and abdominal MDCT. Eur Radiol 24(4):872–880. https://doi.org/10.1007/s00330-013-3089-2
16. Roux C, Rozes A, Reizine D et al (2022) Fully automated opportunistic screening of vertebral fractures and osteoporosis on more than 150 000 routine computed tomography scans. Rheumatol 61(8):3269–3278. https://doi.org/10.1093/rheumatology/keab878

17. Dagan N, Elnekave E, Barda N et al (2020) Automated opportunistic osteoporotic fracture risk assessment using computed tomography scans to aid in FRAX underutilization. Nat Med 26(1):77–82. https://doi.org/10.1038/s41591-019-0720-z
18. Kolanu N, Silverstone EJ, Ho BH et al (2020) Clinical utility of computer-aided diagnosis of vertebral fractures from computed tomography images. J Bone Miner Res 35(12):2307–2312. https://doi.org/10.1002/jbmr.4146
19. Aggarwal V, Maslen C, Abel RL et al (2021) 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 13:1759720X211024029. https://doi.org/10.1177/1759720X211024029
20. Li K, Zhang Y, Wang L et al (2018) The protocol for the Prospective Urban Rural Epidemiology China Action on Spine and Hip status study. Quant Imaging Med Surg 8(7):667–672. https://doi.org/10.21037/qims.2018.08.07
21. He D, Li ZC, Zhang TY, Cheng XG, Tian W (2021) Prevalence of Lumbar Spondylolisthesis in Middle-Aged People in Beijing Community. Orthop Surg 13(1):202–206. https://doi.org/10.1111/os.12871
22. Flahault A, Cadilhac M, Thomas G (2005) Sample size calculation should be performed for design accuracy in diagnostic test studies. J Clin Epidemiol 1;58(8):859-62. https://doi.org/10.1016/j.jclinepi.2004.12.009
23. Nicolaes J, Skjødt MK, Raeymaeckers S et al (2023) Towards improved identification of vertebral fractures in routine CT scans: development and external validation of a machine learning algorithm [in review]
24. Payer C, Stern D, Bischof H, Urschler M (2020) Coarse to Fine Vertebrae Localization and Segmentation with SpatialConfiguration-Net and U-Net. In: Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2020). https://doi.org/10.5220/0008975201240133
25. Sekuboyina A, Husseini ME, Bayat A et al (2021) VerSe: A Vertebrae labelling and segmentation benchmark for multi-detector CT images. Med Image Anal 73:102166. https://doi.org/10.1016/j.media.2021.102166
26. Abadi M, Agarwal A, Barham P, et al (2015) TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org
27. Lowekamp BC, Chen DT, Ibáñez L, Blezek D (2013) The design of SimpleITK. Front Neuroinform Dec 30;7:45
28. 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
29. Link TM (2012) Osteoporosis imaging: state of the art and advanced imaging. Radiol 263(1):3–17. https://doi.org/10.1148/radiol.2633201203
30. Pedregosa F, Varoquaux G, Gramfort A et al (2011) Scikit-learn: Machine learning in Python. J Mach Learn Res 12:2825–2830 https://dl.acm.org/doi/10.5555/1953048.2078195
31. Yushkevich PA, Piven J, Hazlett HC et al (2006) User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. Neuroimage 31(3):1116–1128. https://doi.org/10.1016/j.neuroimage.2006.01.015
32. Ferrar L, Jiang G, Adams J, Eastell R (2005) Identification of vertebral fractures: an update. Osteoporos Int 16(7):717–728. https://doi.org/10.1007/s00198-005-1880-x