Identification of Vertebral Fractures by Convolutional Neural Networks to Predict Nonvertebral and Hip Fractures: A Registry-based Cohort Study of Dual X-ray Absorptiometry

Sheldon Derkatch, MD • Christopher Kirby, MS • Douglas Kimelman, PhD • Mohammad Jafari Jozani, PhD • J. Michael Davidson, MD • William D. Leslie, MD, MSc

From the Department of Radiology, University of Manitoba, 820 Sherbrook St, GA216, Winnipeg, MB, Canada R3T 2N2 (S.D., C.K., D.K., M.J.J., J.M.D., W.D.L.); and St Boniface Hospital Albrechtsen Research Centre, Winnipeg, Canada (C.K., D.K.). Received February 14, 2019; revision requested April 24; revision received July 31; accepted August 5. **Address correspondence to** S.D. (e-mail: *sderkatch@sbgh.mb.ca*).

Conflicts of interest are listed at the end of this article.

Radiology 2019; 00:1–7 • https://doi.org/10.1148/radiol.2019190201 • Content codes: IN MK

Background: Detection of vertebral fractures (VFs) aids in management of osteoporosis and targeting of fracture prevention therapies.

Purpose: To determine whether convolutional neural networks (CNNs) can be trained to identify VFs at VF assessment (VFA) performed with dual-energy x-ray absorptiometry and if VFs identified by CNNs confer a similar prognosis compared with the expert reader reference standard.

Materials and Methods: In this retrospective study, 12 742 routine clinical VFA images obtained from February 2010 to December 2017 and reported as VF present or absent were used for CNN training and testing. All reporting physicians were diagnostic imaging specialists with at least 10 years of experience. Randomly selected training and validation sets were used to produce a CNN ensemble that calculates VF probability. A test set (30%; 3822 images) was used to assess CNN agreement with the human expert reader reference standard and CNN prediction of incident non-VFs. Statistical analyses included area under the receiver operating characteristic curve, two-tailed Student *t* tests, prevalence- and bias-adjusted κ value, Kaplan-Meier curves, and Cox proportional hazard models.

Results: This study included 12 742 patients (mean age, 76 years ± 7; 12 013 women). The CNN ensemble demonstrated an area under the receiver operating characteristic curve of 0.94 (95% confidence interval [CI]: 0.93, 0.95) for VF detection that corresponded to sensitivity of 87.4% (534 of 611), specificity of 88.4% (2838 of 3211), and prevalence- and bias-adjusted κ value of 0.77. On the basis of incident fracture data available for 2813 patients (mean follow up, 3.7 years), hazard ratios adjusted for baseline fracture probability were 1.7 (95% CI: 1.3, 2.2) for CNN versus 1.8 (95% CI: 1.3, 2.3) for expert reader–detected VFs for incident non-VF and 2.3 (95% CI: 1.5, 3.5) versus 2.4 (95% CI: 1.5, 3.7) for incident hip fracture.

Conclusion: Convolutional neural networks can identify vertebral fractures on vertebral fracture assessment images with high accuracy, and these convolutional neural network–identified vertebral fractures predict clinical fracture outcomes.

Online supplemental material is available for this article.

The presence of a previous fracture is a major predictor of future fracture (1,2). Vertebral fractures (VFs) are the most common osteoporosis-related fractures but may not produce acute clinical manifestation (3,4). The radiologic appearances of VF are variable. Moderate and severe fractures have characteristic appearances, but mild VFs can be subtle and be mimicked by nonfracture abnormalities (including Scheuermann disease, degenerative remodeling, scoliosis, and Schmorl nodes) and developmental anomalies (butterfly vertebrae, cupid's bow, and balloon discs) (3–5). The diagnosis of VFs has low interrater reliability at conventional radiography, even in expert hands (6–8). Furthermore, VFs are frequently underdiagnosed in the clinical setting and can be overdiagnosed by using quantitative criteria, speaking to the need for more objective, consistent methods for VF recognition (9).

VF assessment (VFA) is a low-radiation-dose (0.002 to ~0.05 mSv) (10) imaging technique to generate lateral images of the thoracolumbar spine by using dual x-ray absorptiometry (DXA) equipment. VFA is typically performed at the time of bone mineral density (BMD) measurement for the detection of previously undiagnosed VFs because their presence indicates a substantial risk for subsequent fractures independent of BMD. Guidelines suggest the use of targeted VFA to screen for VFs in at-risk populations (11–14).

We hypothesized that, given a sufficiently large training data set that includes a broad spectrum of examples of VF categorized by expert readers, convolutional neural networks (CNNs) could learn to reliably identify VFs. Our aim was to determine if a CNN could be trained to automatically identify the presence of VFs on VFA images as

Abbreviations

BMD = bone mineral density, CI = confidence interval, CNN = convolutional neural network, DXA = dual x-ray absorptiometry, VF = vertebral fracture, VFA = VF assessment

Summary

Deep learning with convolutional neural networks can identify vertebral fractures at lateral imaging of the thoracolumbar spine by using dual x-ray absorptiometry with high accuracy and can predict subsequent nonvertebral and hip fracture outcome as strongly as vertebral fractures diagnosed by human expert readers.

Key Results

- Deep learning with convolutional neural networks (CNNs) had an area under the receiver operating characteristic curve of 0.94 for detection of vertebral fracture (VF), with sensitivity of 87.4% and specificity of 88.4%. The area under the curve was independent of sex, age, body mass index, bone density level, and different scanner generations.
- The CNN had an adjusted κ of 0.76 for agreement with expert reader diagnosis.
- On the basis of incident fracture data for the 2813 patients (mean follow-up, 3.7 years), CNN- and expert reader–detected VFs were similarly predictive of incident fracture. Adjusted hazard ratios for incident non-VF prediction were 1.7 for CNN detected (vs CNN not detected) and 1.8 for expert reader detected (vs not detected) VFs (both $P < .001$); for incident hip fracture prediction, hazard ratios were 2.3 and 2.4, respectively (both $P < .001$).

an aid to fracture risk assessment and if VFs identified by CNNs were similar in prognostic significance when compared with the human expert reader reference standard.

Materials and Methods

Study Cohort

Study patients were residents of Manitoba, Canada, referred for osteoporosis assessment including VFA through the province of Manitoba BMD Program from February 2010 to December 2017. Criteria for referral to the program and for performing VFA are summarized in Tables E1 and E2 (online). This retrospective cohort study was approved by the University of Manitoba Research Ethics Board. The requirement for written informed consent was waived.

All individuals with VFA that was consecutively performed for routine clinical indications during the study period were included in the Manitoba BMD registry and were eligible for analysis (see patient flowchart; Fig 1). Patients were excluded on the basis of expert reader reference standard interpretation of possible or uncertain VF, insufficient image quality for interpretation, or the presence of traumatic or pathologic fracture. If scan height divided by width was less than 0.6, the VFA was considered incomplete and excluded from further analysis. Patients with VFA performed before March 31, 2016, were used for analysis of incident fracture prediction.

VFA Imaging

VFA imaging was performed with first-generation and second-generation fan-beam DXA devices according to manufacturer recommendations (Prodigy or iDXA, respectively; GE Healthcare, Madison, Wis) at the time of BMD measurement. One of four physicians (including W.D.L., a nuclear medicine physician with 31 years of experience reporting BMD) independently assessed the VFA for VFs without any type of blinding. All reporting physicians were diagnostic imaging specialists (radiology and/or nuclear medicine) with at least 10 years of DXA experience and certification by the International Society for Clinical Densitometry. Images were assessed for prevalent VFs by using the modified algorithm-based qualitative method (6). Briefly, vertebrae are considered fractured if there is depression of the vertebral endplate within the vertebral ring (with or without cortical buckling or breaks) in the absence of nonfracture causes of vertebral deformity. Other imaging was used to assist in diagnosis, if available.

To adjust for baseline fracture risk, we calculated 10-year probability of a major osteoporotic fracture including BMD for each study patient by using the commercially available Canadian Frax tool (Frax Desktop Multipatient Entry, version 3.8; Center for Metabolic Bone Diseases, University of Sheffield, England, *www.frax-tool.org*). Fractures detected at VFA were not included in this calculation.

For individuals who underwent VFA imaging before March 31, 2016, longitudinal health service records were assessed to March 31, 2017, for the presence of incident hip fracture and any incident non-VF (including hip but excluding head and neck, hands and feet, and ankle) not associated with codes indicative of severe trauma (ie, external injury) (15).

Anonymized VFA images were resized to 600 × 360 pixels and were randomly assigned to the training (60%; 7646 [of 12742] images), validation (10%; 1274 images), or test set (30%; 3822 images). All images were labeled for the presence or absence of VF as diagnosed by the expert readers in binary format (fracture represented by 1).

An ensemble consisting of an InceptionResNetV2 CNN and a DenseNet CNN was used to help predict the presence of VF (Table E3 [online]). Activation heat maps were automatically generated for each test image, representing the relative significance of areas within a given image to the CNN's VF prediction (16). The code to enable readers to replicate these methods on their own data is available for download online (*https://github.com/DougUC/VFADL-PUBLIC/blob/master/VFADL.ipynb*).

Statistical Analysis

Baseline characteristics of the development and test sets are presented as mean ± standard deviation for continuous variables or number for categorical variables and were compared by two-tailed Student t tests (continuous data) or χ^2 tests (categorical data). Statistical significance was indicated by a P value less than .05. By using the diagnosis of VF as the expert reader reference standard, the performance of the final CNN ensemble output was assessed in the test set for accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve, with 95% confidence intervals (CIs). Overall agreement between the CNN and the original expert reader interpretation of the VFA was also evaluated with the prevalence-adjusted, bias-adjusted κ score (17,18). Kaplan-Meier fracture-free survival curves for any non-VF or hip fracture were compared by using the log-rank

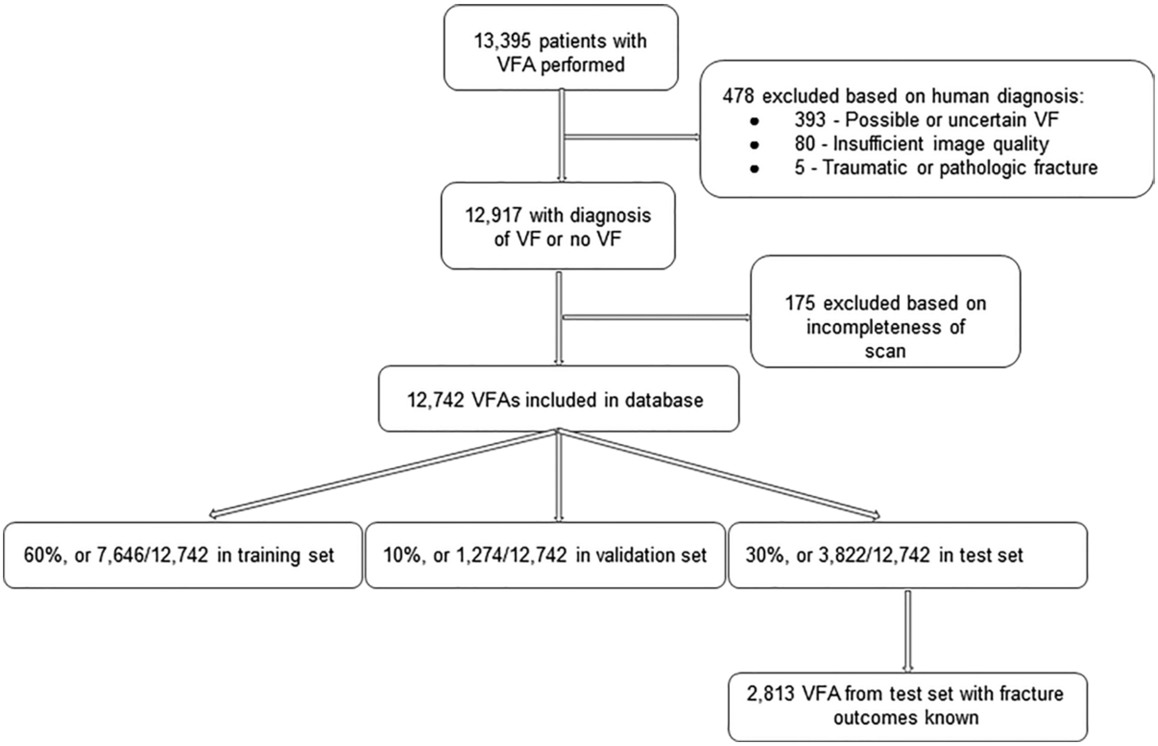
13,395 patients with VFA performed
478 excluded based on human diagnosis:
• 393 - Possible or uncertain VF
• 80 - Insufficient image quality
• 5 - Traumatic or pathologic fracture
12,917 with diagnosis of VF or no VF
175 excluded based on incompleteness of scan
12,742 VFAs included in database
60%, or 7,646/12,742 in training set
10%, or 1,274/12,742 in validation set
30%, or 3,822/12,742 in test set
2,813 VFA from test set with fracture outcomes known

Figure 1: Study patient flowchart. VF = vertebral fracture, VFA = VF assessment.

Table 1: Baseline Characteristics of Test Set Patients with versus without Vertebral Fractures Diagnosed by the Convolutional Neural Network Ensemble and Expert Readers

	CNN			Expert Reader		
Parameter	VF (*n* = 907)	No VF (*n* = 2915)	*P* Value	VF (*n* = 611)	No VF (*n* = 3211)	*P* Value
Mean age (y)	76.9 ± 8.1	75.4 ± 6.7	<.001	76.4 ± 8.7	75.7 ± 6.7	.02
No. of women	823 (90.7)	2773 (95.1)	<.001	548 (89.7)	3048 (94.9)	<.001
Mean body mass index (kg/m^2)	26.0 ± 5.1	26.1 ± 5.1	.73	25.8 ± 5.0	26.2 ± 5.1	.07
Mean lumbar spine T-score	−2.2 ± 1.3	−1.8 ± 1.2	<.001	−2.3 ± 1.2	−1.8 ± 1.2	<.001
Mean femoral neck T-score	−2.0 ± 0.9	−1.7 ± 0.8	<.001	−2.1 ± 0.9	−1.7 ± 0.8	<.001
Mean baseline fracture probability (%)	16.7 ± 7.6	14.5 ± 6.4	<.001	16.8 ± 8.0	14.6 ± 6.4	<.001
No. of 1st-generation DXA scanners	290 (32.0)	946 (32.5)	.79	196 (32.1)	1040 (32.4)	.88

Note.—Mean data are ± standard deviation; data in parentheses are percentages. CNN = convolutional neural network, DXA = dual x-ray absorptiometry, VF = vertebral fracture.

test according to CNN and expert reader diagnosis of prevalent (baseline) VF detected at VFA. Cox proportional hazards models were used to generate hazard ratios with 95% CIs for time to first non-VF and time to first hip fracture according to CNN and expert reader diagnosis of VFs detected at VFA. Statistical analyses were performed by using scikit-learn (19) (Statistica version 13.0; Tibco Software, Palo Alto, Calif) and SPSS for Windows (version 24; IBM, Armonk, NY).

Results

The final analytical data set was composed of 12742 VFA images from unique patients (mean age, 76 years ± 7; 12013 women), among which 2116 (16.6%) had one or more expert reader–diagnosed VFs (Fig 1). The training and validation sets were composed of a development set of 8920 VFAs from unique patients (mean age, 76 years ± 7; 94.4% women [8417 of 8929]). The test set consisted of 3822 VFAs from unique patients (mean age, 76 years ± 7; 94.1% women [3596 of 3822]). The development and test image sets did not differ in baseline characteristics, including age (*P* = .49), sex (*P* = .54), body mass index (*P* = .73), lumbar spine T-score (*P* = .94), femoral neck T-score (*P* = .29), baseline fracture probability (*P* = .59), or scanner generation (*P* = .47) (Table E4 [online]).

For both the CNN ensemble and expert readers, patients from the test set diagnosed with VFs were older than those without VFs (CNN: mean age, 76.9 years ± 8.1 vs 75.4 years ± 6.7, respectively, *P* < .001; expert readers: 76.4 years ± 8.7 vs 75.7 years ± 6.7, respectively, *P* = .02), had lower BMD T-scores (lumbar T-scores, CNN: −2.2 ± 1.3 vs −1.8 ± 1.2, respectively, *P* < .001; lumbar T-scores, expert readers: −2.3 ± 1.2 vs

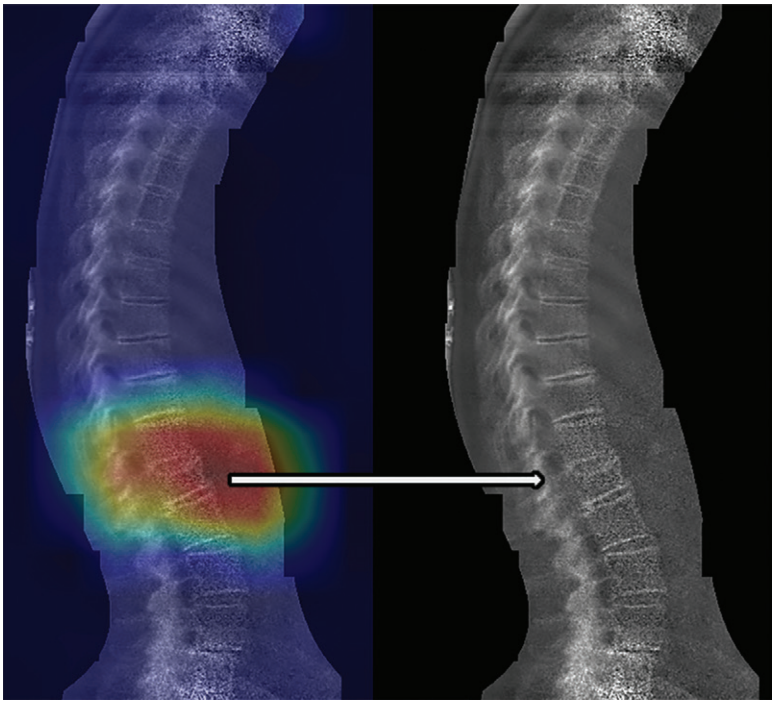

Figure 2: Images in a 72-year-old female patient evaluated for vertebral fracture (VF). VF heat map shows a mild vertebral compression fracture. Heat maps are unitless low-resolution images generated from high-level activations and gradients within the convolutional neural network (CNN). They have been magnified and overlaid on the original image to show the relative importance of large regions of the original image to the ultimate CNN prediction. The heat map has been overlaid on the original VF assessment image (left side). The arrow shows the corresponding location of VF on the original image, as presented to the convolutional neural network (right side).

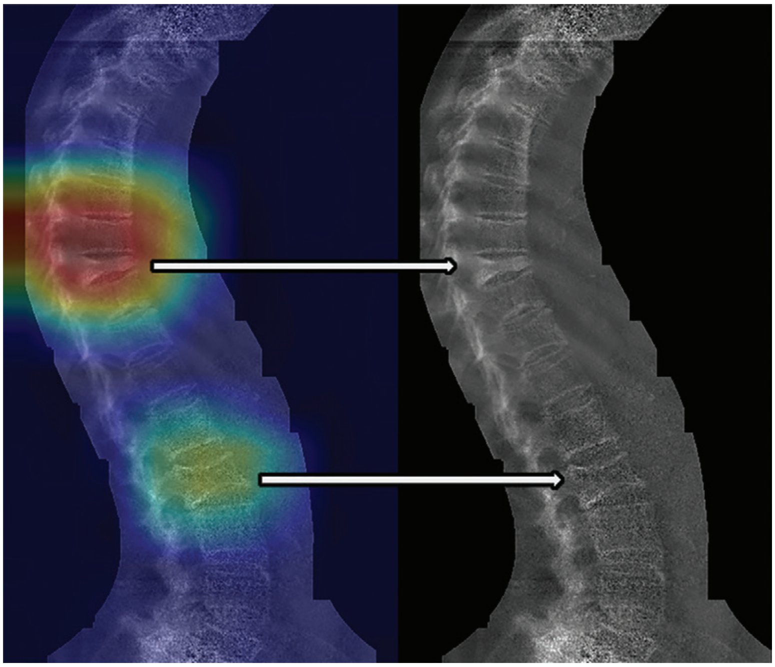

Figure 3: Images in a 77-year-old female patient evaluated for vertebral fracture (VF). VF heat map shows one severe vertebral compression fracture (upper arrow) and one mild fracture (lower arrow). Heat maps are unitless low-resolution images showing relative contributions of general areas in images to the prediction. The heat map has been overlaid on the original VF assessment image (left side). Arrows denote the corresponding locations of VFs on the original images, as presented to the convolutional neural network (right side).

−1.8 ± 1.2, respectively, $P < .001$; femoral neck T-scores, CNN: −2.0 ± 0.9 vs −1.7 ± 0.8, $P < .001$; femoral neck T-scores, expert readers: −2.1 ± 0.9 vs −1.7 ± 0.8, respectively, $P < .001$), and had higher baseline fracture probability calculated without inclusion of the VFA results (CNN, 16.7% ± 7.6 vs 14.5% ± 6.4, respectively, $P < .001$; expert readers, 16.8% ± 8.0 vs 14.6% ± 6.4, respectively, $P < .001$) (Table 1). Unitless heat maps of relative regional significance to the CNN showed fracture localization capability (Figs 2, 3).

The CNN ensemble had an accuracy of 3372 of 3822 (88.2%; 95% CI: 87.2%, 89.2%) for VF detection compared with expert reader diagnosis. Sensitivity and specificity were 534 of 611 (87.4%; 95% CI: 86.3%, 88.5%) and 2838 of 3211 (88.4%; 95% CI: 87.4%, 89.4%), respectively. The prevalence- and bias-adjusted κ score for agreement with expert reader diagnosis was 0.77 (95% CI: 0.74, 0.79). On the basis of the continuous CNN output, the area under the receiver operating characteristic curve was 0.94 (95% CI: 0.93, 0.95) for VF detection. No differences in performance were apparent when stratified by sex (P = .16), age (P = .88), obesity (P = .34), osteoporotic BMD T-score (P = .76), and scanner type (P = .94) (Table 2). Areas under the receiver operating characteristic curve for the component CNNs are provided in Table E5 (online).

Of the 3822 individuals in the test group, 2813 were imaged before March 31, 2016, and were assessed for incident fractures. During mean 3.7 years, 85 (3.0%) patients had incident hip fractures and 247 (8.8%) had incident non-VFs. Fracture-free survival was significantly worse after diagnosis of VF, whether by the CNN ensemble or by expert readers (all $P < .001$) (Fig 4). Hazard ratios for hip fracture and any non-VF were higher after diagnosis of VF by the CNN ensemble or by expert readers (Table 3), both before and after adjustment for baseline fracture probability ($P < .001$ for each). Adjusted hazard ratios for hip fracture for CNN and expert readers were 2.3 (95% CI: 1.5, 3.5) and 2.4 (95% CI: 1.5, 3.7), respectively. The adjusted hazard ratios for any non-VF were 1.7 (95% CI: 1.3, 2.2) and 1.8 (95% CI: 1.3, 2.3). Results were unchanged when further adjusted for osteoporosis treatment. Hazard ratios were also calculated for patients grouped by expert reader–CNN agreement (Table E6 [online]). Agreement for VF (both CNN and expert reader positive for VF vs both negative for VF) was associated with slightly higher adjusted hazard ratios for any non-VF (hazard ratio, 2.0; 95% CI: 1.5, 2.7) or hip fracture (hazard ratio, 2.8; 95% CI: 1.7, 4.4). When the CNN diagnosed a VF but the expert reader did not, there was a higher risk of incident fracture but this was not statistically

Table 2: Convolutional Neural Network Ensemble Performance for Vertebral Fracture Diagnosis versus Human Readers, Overall and Stratified by Subgroups

Parameter	No. of Patients	Sensitivity (%)	Specificity (%)	Accuracy (%)	AUC	*P* Value
Overall	3822	87.4 (86.3, 88.5)	88.4 (87.4, 89.4)	88.2 (87.2, 89.2)	0.94 (0.93, 0.95)	
Patients older than median age of 76 years	1889	89.2 (87.8, 90.6)	86.1 (84.5, 87.7)	86.7 (85.2, 88.2)	0.94 (0.93, 0.96)	0.88
Patients younger than median age of 76 years	1933	84.9 (83.3, 86.5)	90.4 (89.1, 91.7)	89.7 (88.3, 91.1)	0.94 (0.92, 0.96)	…
Men	226	84.1 (79.3, 88.9)	81.0 (75.9, 86.1)	81.9 (76.9, 86.9)	0.91 (0.86, 0.96)	0.16
Women	3596	87.8 (86.7, 89.9)	88.8 (87.8, 89.8)	88.6 (87.6, 89.6)	0.94 (0.93, 0.96)	…
BMI ≥ 30 kg/m^2	751	83.6 (81.0, 86.2)	86.5 (84.1, 88.9)	86.0 (83.5, 88.5)	0.93 (0.90, 0.96)	0.34
BMI < 30 kg/m^2	3071	88.3 (87.2, 89.4)	88.9 (87.8, 90.0)	88.8 (87.7, 89.9)	0.94 (0.93, 0.96)	…
Osteoporotic	1023	90.9 (89.1, 92.7)	84.5 (82.3, 86.7)	86.0 (83.9, 88.1)	0.94 (0.92, 0.96)	0.76
Nonosteoporotic	2799	85.1 (83.8, 86.4)	89.6 (88.5, 90.7)	89.0 (87.8, 90.2)	0.94 (0.92, 0.95)	…
First-generation DXA scanner	1236	86.2 (84.3, 88.1)	88.4 (86.6, 90.2)	88.0 (86.2, 89.8)	0.94 (0.92, 0.96)	0.94
Second-generation DXA scanner	2586	88.0 (86.7, 89.3)	88.0 (87.2, 89.6)	88.3 (87.1, 89.5)	0.94 (0.93, 0.96)	…

Note.—Data in parentheses are 95% confidence intervals. AUC = area under the receiver operating characteristic curve, BMI = body mass index, DXA = dual x-ray absorptiometry.

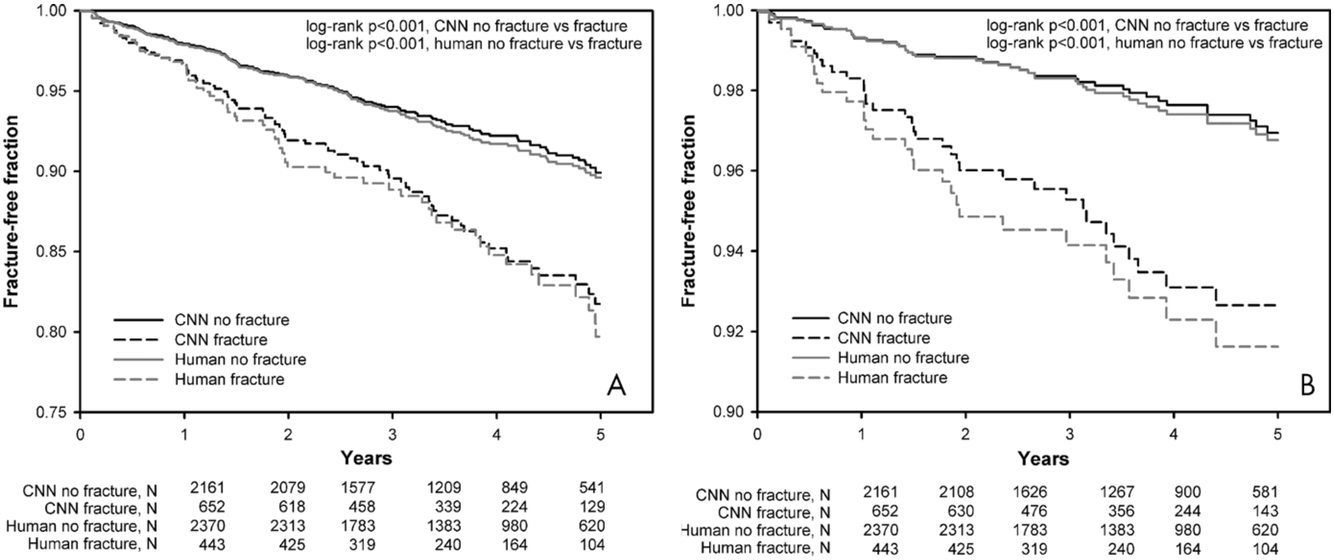

Figure 4: Fracture-free survival curves of, *A*, any non-VF and, *B*, hip fractures according to convolutional neural network (CNN) ensemble and expert reader diagnosis of prevalent (baseline) vertebral fracture (VF) at VF assessment.

significant (adjusted hazard ratio: any non-VF, 1.3 [95% CI: 0.8, 1.9], *P* = .28; hip fracture: 1.5 [95% CI: 0.8, 3.0], *P* = .26), whereas no higher risk was observed when the expert reader diagnosed a VF and the CNN did not.

Discussion

Vertebral fractures (VFs) can be a diagnostic challenge because of variability in clinical manifestation and overlap in appearances between mild VF and vertebral configurations not from fractures. We sought to determine if convolutional neural networks (CNNs) could be trained to automatically identify VFs on VF assessment (VFA) images acquired by using dual x-ray absorptiometry (DXA) with only binary diagnostic labels, and to assess the prognostic significance of fractures identified by CNNs compared with the expert reader reference standard. We found that our CNN ensemble could detect VFs with a sensitivity of 87.4% (95% confidence interval [CI]: 86.3%, 88.5%), specificity of 88.4% (95% CI: 87.4%, 89.2%), and area under the receiver operating characteristic curve of 0.94 (95% CI: 0.93, 0.95), and that these fractures have clinical significance similar to those detected by the expert reader reference standard, with adjusted hazard ratios for future hip fracture of 2.3 (95% CI: 1.5, 3.5) for CNNs versus 2.4 (95% CI: 1.5, 3.8) for expert readers (both $P < .001$). Our CNN ensemble performed equally well within subgroups defined by age (*P* = .88), sex (*P* = .16), body mass index (*P* = .34), bone mineral density (BMD) T-score (*P* =

Table 3: Incident Fracture for Groups with and without Diagnosis of Vertebral Fracture by Convolutional Neural Network Ensemble and Expert Readers

Parameter	Model 1: Unadjusted Hazard Ratio	Model 2: Adjusted for Baseline Fracture Probability Hazard Ratio	Model 3: Adjusted for Baseline Fracture Probability and Osteoporosis Treatment Hazard Ratio
Outcome, any non-VF*			
VF vs No VF with CNN	2.0 (1.5, 2.6)	1.7 (1.3, 2.2)	1.7 (1.3, 2.2)
VF vs No VF with expert reader	2.1 (1.5, 2.7)	1.8 (1.3, 2.3)	1.8 (1.3, 2.4)
Outcome, hip fracture			
VF vs No VF with CNN	3.0 (2.0, 4.6)	2.3 (1.5, 3.5)	2.3 (1.5, 3.6)
VF vs No VF with expert reader	3.1 (2.0, 4.8)	2.4 (1.5, 3.7)	2.5 (1.6, 3.9)

Note.—Data are from Cox proportional hazards models; data in parentheses are 95% confidence intervals. CNN = convolutional neural network, VF = vertebral fracture.

* Excludes head and neck, hands and feet, and ankle.

.76), and generation of DXA scanner (*P* = .94). Agreement between CNN and expert reader reference standard diagnosis was substantial (prevalence- and bias-adjusted κ, 0.77).

To our knowledge, this is the first attempt to use CNNs to identify VFs on VFA images (20). Our level of CNN-to–expert reader agreement is similar to and often better than the level of expert reader–to–expert reader interobserver agreement seen with VFs diagnosed on radiographs (7). In a large population-based study (7582 patients), the prevalence- and bias-adjusted κ for VF diagnosis at radiography was 0.74 overall but lower in older individuals (κ, 0.68, age 70–79 years; and κ, 0.66, age ≥80 years) (7). It is also worth noting the potential advantage of access to ancillary diagnostic imaging available to the expert readers in our study, which was not available to the CNNs, that might affect agreement in some individuals.

Strengths of our study include the use of large high-quality training, validation, and test data sets; relevant covariates; and follow-up data on fracture outcomes. The expert reader reference standard VFA interpretations were from a small number of experienced clinical readers who used validated criteria for VF diagnosis and (when available) ancillary imaging data. Although not on the basis of a random sample, our results are likely to be representative of routine clinical practice and the older population typically referred for bone densitometry.

One potential limitation of our study was the use of ancillary imaging data in clinical reference standard diagnosis of VF. This additional data may have adversely affected CNN training and evaluation relative to their actual informational content. However, the use of such external data would be expected to improve the accuracy of the expert reader reference standard and result in improved CNN training and generalization. Our study cohort contained a relatively small proportion of men, consistent with expected BMD referral patterns, limiting our confidence in that subgroup.

Our work shows that an entirely automated method of vertebral fracture (VF) detection can predict clinically relevant fracture outcomes with an accuracy similar to expert readers while avoiding expert reader intraobserver and interobserver variability. This technology may enhance the diagnostic performance of radiologists and nuclear medicine physicians. It could improve consistency of reporting of incidentally detected VFs, leading to earlier therapeutic intervention and better patient outcomes. More consistent VF recognition could also facilitate the wider incorporation of VF assessment into clinical practice and lead to more standardized diagnosis in osteoporosis research. Assessing the generalizability and transferability of our convolutional neural networks (CNNs) to other populations and scanning equipment (eg, single-energy VF assessment, radiography) is an important next step. There may be physician and patient resistance to the use of inscrutable so-called black box machine interpretation, an area that requires further study. We illustrate one approach to support the validity of the CNN diagnoses by providing insight into what the algorithm is responding to on the basis of activation heat maps (30). An image segmentation approach to CNN training and evaluation could further improve both fracture detection and stakeholder acceptance. Such an approach would require expert image interpreters to label regions of interest. Further assessment of the practical use of this technology will require integration into clinical workflows for prospective validation, and acceptance by stakeholders including hospital administrators, physicians (both radiologists and nonradiologists), and patients.

Acknowledgments: The authors acknowledge the Manitoba Centre for Health Policy for use of data contained in the Population Health Research Data Repository (HIPC 2016/2017–29). The results and conclusions are those of the authors and no official endorsement by the Manitoba Centre for Health Policy, Manitoba Health, Seniors and Active Living, or other data providers is intended or should be inferred. This article has been reviewed and approved by the members of the Manitoba Bone Density Program Committee. S.D. acknowledges the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

Author contributions: Guarantors of integrity of entire study, S.D., W.D.L.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, S.D., D.K., M.J.J., W.D.L.; clinical studies, S.D., J.M.D., W.D.L.; experimental studies, S.D., C.K., D.K.; statistical analysis, S.D., C.K., D.K., M.J.J., W.D.L.; and manuscript editing, S.D., C.K., D.K., M.J.J., W.D.L.

Disclosures of Conflicts of Interest: **S.D.** disclosed no relevant relationships. **C.K.** disclosed no relevant relationships. **D.K.** disclosed no relevant relationships. **M.J.J.** disclosed no relevant relationships. **J.M.D.** disclosed no relevant relationships. **W.D.L.** disclosed no relevant relationships.

References

1. Kanis JA, Johnell O, De Laet C, et al. A meta-analysis of previous fracture and subsequent fracture risk. Bone 2004;35(2):375–382.
2. Klotzbuecher CM, Ross PD, Landsman PB, Abbott TA 3rd, Berger M. Patients with prior fractures have an increased risk of future fractures: a summary of the literature and statistical synthesis. J Bone Miner Res 2000;15(4):721–739.
3. Jiang G, Eastell R, Barrington NA, Ferrar L. Comparison of methods for the visual identification of prevalent vertebral fracture in osteoporosis. Osteoporos Int 2004;15(11):887–896.
4. Oei L, Rivadeneira F, Ly F, et al. Review of radiological scoring methods of osteoporotic vertebral fractures for clinical and research settings. Eur Radiol 2013;23(2):476–486.
5. Lentle B, Koromani F, Brown JP, et al. The Radiology of Osteoporotic Vertebral Fractures Revisited. J Bone Miner Res 2019;34(3):409–418.
6. Lentle BC, Berger C, Probyn L, et al. Comparative Analysis of the Radiology of Osteoporotic Vertebral Fractures in Women and Men: Cross-Sectional and Longitudinal Observations from the Canadian Multicentre Osteoporosis Study (CaMos). J Bone Miner Res 2018;33(4):569–579.
7. Oei L, Koromani F, Breda SJ, et al. Osteoporotic Vertebral Fracture Prevalence Varies Widely Between Qualitative and Quantitative Radiological Assessment Methods: The Rotterdam Study. J Bone Miner Res 2018;33(4):560–568.
8. Aubry-Rozier B, Chapurlat R, Duboeuf F, et al. Reproducibility of Vertebral Fracture Assessment Readings From Dual-energy X-ray Absorptiometry in Both a Population-based and Clinical Cohort: Cohen's and Uniform Kappa. J Clin Densitom 2015;18(2):233–238.
9. Szulc P. Vertebral Fracture: Diagnostic Difficulties of a Major Medical Problem. J Bone Miner Res 2018;33(4):553–559.
10. Damilakis J, Adams JE, Guglielmi G, Link TM. Radiation exposure in X-ray-based imaging techniques used in osteoporosis. Eur Radiol 2010;20(11):2707–2714.
11. Cosman F, de Beur SJ, LeBoff MS, et al. Clinician's Guide to Prevention and Treatment of Osteoporosis. Osteoporos Int 2014;25(10):2359–2381 [Published correction appears in Osteoporos Int 2015;26(7):2045–2047.] https://doi.org/10.1007/s00198-014-2794-2.
12. Schousboe JT, Vokes T, Broy SB, et al. Vertebral Fracture Assessment: the 2007 ISCD Official Positions. J Clin Densitom 2008;11(1):92–108.
13. Papaioannou A, Morin S, Cheung AM, et al. 2010 clinical practice guidelines for the diagnosis and management of osteoporosis in Canada: summary. CMAJ 2010;182(17):1864–1873.
14. Compston J, Cooper A, Cooper C, et al. UK clinical guideline for the prevention and treatment of osteoporosis. Arch Osteoporos 2017;12(1):43.
15. Lix LM, Azimaee M, Osman BA, et al. Osteoporosis-related fracture case definitions for population-based administrative data. BMC Public Health 2012;12(1):301.
16. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. See https://arxiv.org/abs/1610.02391 v3. 2016;7(8). https://github.com/ramprs/grad-cam/. Accessed August 27, 2019.
17. Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. J Clin Epidemiol 1993;46(5):423–429.
18. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33(1):159–174.
19. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res 2011;12(Oct):2825–2830. http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf.
20. 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.