Validation of a Deep Learning Tool for Detection of Incidental Vertebral Compression Fractures

Michelle Dai, BSc,† Bryan-Clement Tiu, BSc, MPH,* Jacob Schlossman, BSc,* Angela Ayobi, MSc, MEng,‡ Charlotte Castineira, MEng,‡ Julie Kiewsky, MEng,‡ Christophe Avare, PhD,‡ Yasmina Chaibi, PhD,‡ Peter Chang, MD,§‖ Daniel Chow, MD,§‖ and Jennifer E. Soun, MD§*

Objective: This study evaluated the performance of a deep learning-based vertebral compression fracture (VCF) detection tool in patients with incidental VCF. The purpose of this study was to validate this tool across multiple sites and multiple vendors.

Methods: This was a retrospective, multicenter, multinational blinded study using anonymized chest and abdominal CT scans performed for indications other than VCF in patients ≥ 50 years old. Images were obtained from 2 teleradiology companies in France and United States and were processed by CINA-VCF v1.0, a deep learning algorithm designed for VCF detection. Ground truth was established by majority consensus across 3 board-certified radiologists. Overall performance of CINA-VCF was evaluated, as well as subset analyses based on imaging acquisition parameters, baseline patient characteristics, and VCF severity. A subgroup was also analyzed and compared with available clinical radiology reports.

Results: Four hundred seventy-four CT scans were included in this study, comprising 166 (35.0%) positive and 308 (65.0%) negative VCF cases. CINA-VCF demonstrated an area under the curve (AUC) of 0.97 (95% CI: 0.96-0.99), accuracy of 93.7% (95% CI: 91.1%-95.7%), sensitivity of 95.2% (95% CI: 90.7%-97.9%), and specificity of 92.9% (95% CI: 89.4%-96.5%). Subset analysis based on VCF severity resulted in a specificity of 94.2% (95% CI: 90.9%-96.6%) for grade 0 negative cases and a specificity of 64.3% (95% CI: 35.1%-87.2%) for grade 1 negative cases. For grades 2 and 3 positive cases, sensitivity was 89.7% (95% CI: 79.9%-95.8%) and 99.0% (95% CI: 94.4%-100.0%), respectively.

Conclusions: CINA-VCF successfully detected incidental VCF and even outperformed clinical reports. The performance was consistent among all subgroups analyzed. Limitations of the tool included various confounding pathologies such as Schmorl's nodes and borderline cases. Despite these limitations, this study validates the applicability and generalizability of the tool in the clinical setting.

Key Words: compression fracture, spinal fracture, computed tomography, deep learning

(*J Comput Assist Tomogr* 2025;49:669–674)

Vertebral compression fractures (VCFs) are the most common osteoporotic fractures, with an estimated prevalence of 5.3 million cases worldwide.[1] VCFs are associated with increased morbidity, leading to chronic back pain, progressive kyphosis, and impaired pulmonary function.[2] Thus, prompt diagnosis of VCFs is critical to guide appropriate management, prevent disability, and decrease future fracture risk, particularly in the elderly population.

Many VCFs are discovered incidentally on CT of the chest, abdomen, or pelvis obtained for indications unrelated to the VCF. However, VCFs are often underreported by radiologists.[3–7] False negative rates have been reported as high as 46.5% in one multicenter, multinational study.[8]

To help mitigate this issue, deep learning tools have been developed for VCF detection.[9–11] These tools provide automatic results, potentially reducing the workload and improving the accuracy of radiologists screening for incidental findings on a busy worklist. However, the utility of these tools may be limited in real-world clinical settings.[12] Given the paucity of data on the generalizability of automated VCF detection, this study aims to evaluate the performance of a deep learning-based FDA-cleared VCF detection tool, CINA-VCF v1.0 (Avicenna.AI, La Ciotat, France), in clinical settings across multiple sites and multiple vendors in the United States and France.

MATERIALS AND METHODS

This was a retrospective study validating CINA-VCF for VCF detection. Author J.E.S. controlled the data for the study and has no commercial or financial relationships that could be construed as a potential conflict of interest. The data set was sourced from 2 teleradiology networks: vRAD (Minneapolis, MN) and TeleDiag (Lyon, France). A waiver of consent was obtained from the Western IRB and TeleDiag for vRAD and TeleDiag cases, respectively. Informed consent for participation was not required for this study in accordance with the national legislation and

From the *Irvine School of Medicine, University of California, Irvine, CA; †Touro University Nevada, College of Osteopathic Medicine, Henderson, NV; ‡Avicenna.AI, La Ciotat, France; §Department of Radiological Sciences; and ‖Center for Artificial Intelligence in Diagnostic Medicine, University of California Irvine, Irvine, CA.

Received for publication October 8, 2024; accepted December 7, 2024.

Correspondence to: Jennifer E. Soun, MD, Irvine School of Medicine, University of California, Orange 92868, CA (e-mail: jesoun@uci.edu.

M.D. and B.T. are equal contributors to this work and are designated as co-first authors.

A.A., C.C., J.K., C.A., and Y.C. are employees of Avicenna.AI. P.C. is co-founder of and owns stock in Avicenna.AI, has past and current research funding from and is a paid consultant for Canon Medical, has a current grant from Novocure, has past research funding from GE, is a paid consultant for Bayer, and is a paid consultant and speaker for Siemens. D.C. owns stock in Avicenna.ai and is a paid consultant for Canon Medical. J.E.S. controlled the data for the study. The remaining authors declare no conflict of interest.

Supplemental Digital Content is available for this article. Direct URL citations are provided in the HTML and PDF versions of this article on the journal's website, www.jcat.org.

DOI: 10.1097/RCT.0000000000001726

institutional requirements. The data were anonymized except for study date, allowing for partition into non-overlapping training and validation data sets.

Patient Selection

Inclusion in this study required cases to be performed on adults 50 years of age or older and for nonenhanced or contrast-enhanced CT scans to meet the following criteria: field of view includes the chest and/or abdomen, axial or sagittal acquisition with a homogeneous slice interval without a gap between successive slices, slice thickness ≤ 3 mm and soft and standard tissue reconstruction kernel.

Exclusion criteria were CT scans performed on patients under 50 years of age or scans not following the mandatory acquisition parameters listed above, as well as scans with the following parameters: only the cervical spine was visible, at least 3 consecutive vertebrae in the T1-L5 portion of the spine were not present and not completely visible, uninterpretable poor-quality images (ie, significant motion, streak, or other artifacts impeding CT interpretation), and redundant cases.

Deep Learning Algorithm

CINA-VCF is a deep learning algorithm that employs a cascade of task-focused networks based on 2D and 3D U-Nets derived from convolutional neural networks (CNNs).[13] First, the algorithm locates the spine to reduce data size and standardize the field of view. Subsequently, it identifies the center of each lumbar and thoracic vertebral body. A postprocessing step filters out artifacts such as cement or other materials. The algorithm then identifies 6 measurement landmarks per vertebra to calculate vertebral height loss (VHL). These values are then used to passively flag a case as positive or negative for VCF.

The algorithm was trained and tested on a data set representing 12,402 vertebrae acquired between January and March 2021 for vRAD and March 2022 for TeleDiag using a 5-fold approach (80% training and 20% testing) to mitigate training bias. During the training phase, data augmentation was used to enhance the data sets and to improve model optimization and generalizability.

Subsequently, a sensitivity analysis was performed on a separate data set of 1994 vertebrae from 152 cases to set the internal thresholds and achieve a statistical power of 80% at the 5% risk level for the triage metrics 95% CI lower bound. The cases were acquired by vRAD during March 2021 and by TeleDiag during June 2022. The resulting performance, yielding a sensitivity of 92% (95% CI: 82%-97%), specificity of 99% (95% CI: 93%-100%), and accuracy of 96% (95% CI: 92%-99%), demonstrated sufficient robustness of the models, allowing for final clinical validation to proceed.

The validation data set was consecutively acquired between April and November 2021 for vRAD and between November 2022 and January 2023 for TeleDiag. All data sets were adequately distributed across different vendors, patient demographics (age and gender), presence of contrast, fields of view, and slice thicknesses. Both thoracic and lumbar vertebrae were uniformly represented, and there was substantial variation in VHL.

VCF Assessment

Ground truth was established by 2 board-certified neuroradiologists (J.E.S. and D.C.), with consensus for discordant cases determined by a third board-certified neuroradiologist (P.C.). The neuroradiologists were allowed to use measurement tools to proceed with the VHL measurement in indeterminate cases and report any confounding conditions (ie, artifact, presence of bone lesion, and Schmorl's nodes).

VCFs were graded by the Genant method, which assigns to each vertebra a specific grade (normal, mild, moderate, and severe) based on the degree of VHL.[14] VHL of $< 20\%$ reduction was graded as normal (grade 0), 20% to 25% reduction with mild but "definite" fracture as mild (grade 1), 25% to 40% reduction as moderate (grade 2), and 40% or greater reduction as severe (grade 3). Normal and mild fracture grades were grouped together and defined as negative, similar to previous studies and in accordance with FDA clearance.[15,16]

Statistical Analysis

Interrater reliability of ground truth assessment was evaluated by Cohen kappa coefficient. Predictions from the CINA-VCF device were compared with ground truth through a confusion matrix to obtain sensitivity, specificity, and accuracy. Positive predictive values (PPV) and negative predictive values (NPV) were computed with varying prevalence values (from 5% to 50%, in increments of 5%). Stratified analyses were also performed on scanning parameters, patient baseline characteristics (geographical distribution of scans, age, sex, and ethnicity), and Genant severity grading. A subgroup was also compared with clinical radiology reports. The minimum number of cases needed to achieve an expected area under the receiver operating characteristic curve (ROC-AUC) at 95%, with a significance of 5% and a power of 80% ($\beta = 0.2$), were 133 positive and 133 negative cases. MedCalc Statistical Software (v20.015; MedCalc Software Ltd, Ostend, Belgium) was used for analyses.

RESULTS

In total, 491 CT exams (331 vRAD and 160 TeleDiag) met inclusion criteria. Seventeen (3.5%) cases were excluded due to discordance among the 3 ground truth experts or indeterminate diagnosis. A total of 474 anonymized CT cases (259 with contrast and 215 without contrast), including 317 US vRAD cases and 157 TeleDiag cases, were included at this stage for CINA-VCF standalone performance validation. Confounding conditions, including artifact (motion, streak, or noise), was present in 9 positive and 16 negative cases; bone lesions (sclerotic or lytic foci, hemangioma, or metastasis) in 13 positive cases and 19 negative cases; and Schmorl's nodes in 63 positive cases and 102 negative cases. Some cases had more than one confounding condition. The mean $\pm$ SD age of patients included in the study was 72.1 ± 10.1 years old. Patient sex distribution was 50.8% female, 47.7% male, and 1.5% where the information was not available. Thirty-eight different scanner models were utilized: 13, 8, 12, and 5 different models for GE (Chicago, IL), Philips (Amsterdam, Netherlands), Siemens (Munich, Germany), and Canon (Toshiba; Otawara, Tochigi, Japan), respectively. There were 294 (62.0%) grade 0 cases, 14 (3.0%) grade 1 cases, 68 (14.3%) grade 2 cases, and 98 (20.7%) grade 3 cases. Patient characteristics and CT scan parameters are presented in Supplementary Table 1, Supplemental Digital Content 1, http://links.lww.com/RCT/A352.

Among the 474 cases, 1 US board-certified radiologist (J.S.), assessed 155 cases as positive (ie, at least 1 vertebra

presented a VCF of grade 2 or more) and 312 as negative. The second US board-certified radiologist (D.C.) annotated 154 cases as positive and 314 as negative. Disagreements were observed for 51 (10.8%) cases (Cohen Kappa = 0.75, 95% CI: 0.69-0.81). The third US board-certified radiologist (P.C.) reviewed the discordant cases, and ground truth was established by majority consensus. The final VCF validation data set included 166 (35.0%) positive and 308 (65.0%) negative cases. Figure 1 shows examples of true positive (TP), false negative (FN), and false positive (FP) cases.

The ROC-AUC was 0.974 (95% CI: 0.962-0.986) (Fig. 2). The algorithm demonstrated sensitivity of 95.2% (90.7%-97.9%), specificity of 92.9% (89.4%-96.5%), and accuracy of 93.7% (95% CI: 91.1%-95.7%) (Fig. 3). There were 8 FN with a miss rate of VCF of 4.8% (8/166 positive cases). Most of the FN cases (5/8) were mild or borderline VCFs with VHL around 25% (Fig. 1B). Other FNs were caused by the concomitance of VCF with other pathologies or anatomic distortions such as Schmorl's nodes (5/8), scoliosis (2/8), or artifact (2/8). A total of 22/308 (7.1%) cases were FP. This was caused by Schmorl's nodes (Fig. 1C) and other mimics of VCF.

Subset performance analysis was performed on a per-case basis according to Genant grade (Figs. 4, 5). Please note that the per-case VCF grade was defined according to the vertebra with the highest VHL grading. For example, grade 3 VCF cases may have also included vertebrae with grade 2 and/or grade 1 VCF. For grade 0 negative cases, specificity was 94.2% (95% CI: 90.9%-96.6%), and for grade 1 negative cases, specificity was 64.3% (95% CI: 35.1%-87.2%). For grades 2 and 3 positive cases, sensitivity was 89.7% (95% CI: 79.9%-95.8%) and 99.0% (95% CI: 94.4%-100.0%), respectively. Subset performance analyses for imaging parameters (scanner manufacturers, slice thickness, contrast presence, and reconstruction kernel) and patient baseline characteristics (age and sex) ranged from 91.6% to 100%, and specificities from 97.3% to 100%.

The clinical radiology report was available for 317 cases from the total cohort of 474 cases. Among these 317 cases, there were 120 positive cases according to the ground truth. Only 44 of these cases were mentioned as positive in the clinical radiology report; for the remaining 76 cases, there was no mention of VCF in the report, leading to a missed rate of 63.3% (76/120). CINA-VCF was capable of correctly detecting 111 of the 120 positive cases, yielding a missed rate of 7.5% (9/120). Therefore, the algorithm was capable of detecting 88.2% (67/76) of cases that were initially missed or not mentioned by the clinical report.

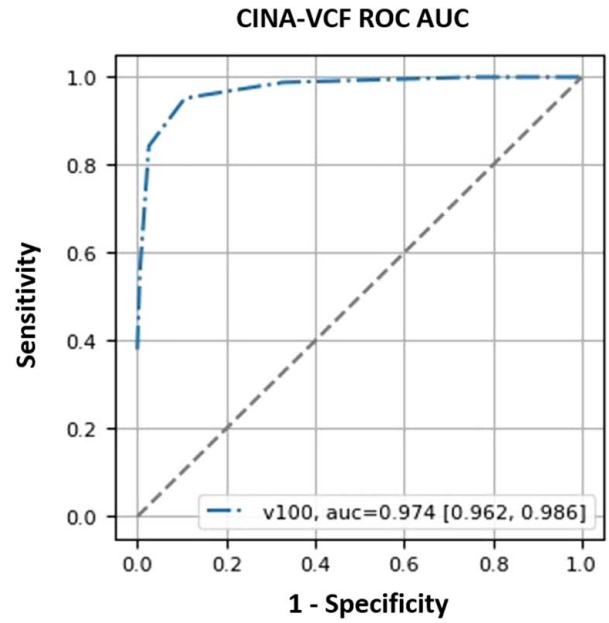

FIGURE 2. Area under the receiver operating characteristic curve (ROC-AUC) for CINA-VCF.

DISCUSSION

This is the first study to validate the performance of CINA-VCF in detecting incidental VCFs in the clinical setting. CINA-VCF demonstrated an area under the curve (AUC) of 0.97, accuracy of 93.7%, sensitivity of 95.2%, and specificity of 92.9%. When accounting for VCF severity, CINA-VCF had a specificity of 94.2% for grade 0 negative cases and 64.3% for grade 1 negative cases. For grades 2 and 3 positive cases, sensitivity was 89.7% and 99.0%, respectively. Performance remained consistent across imaging

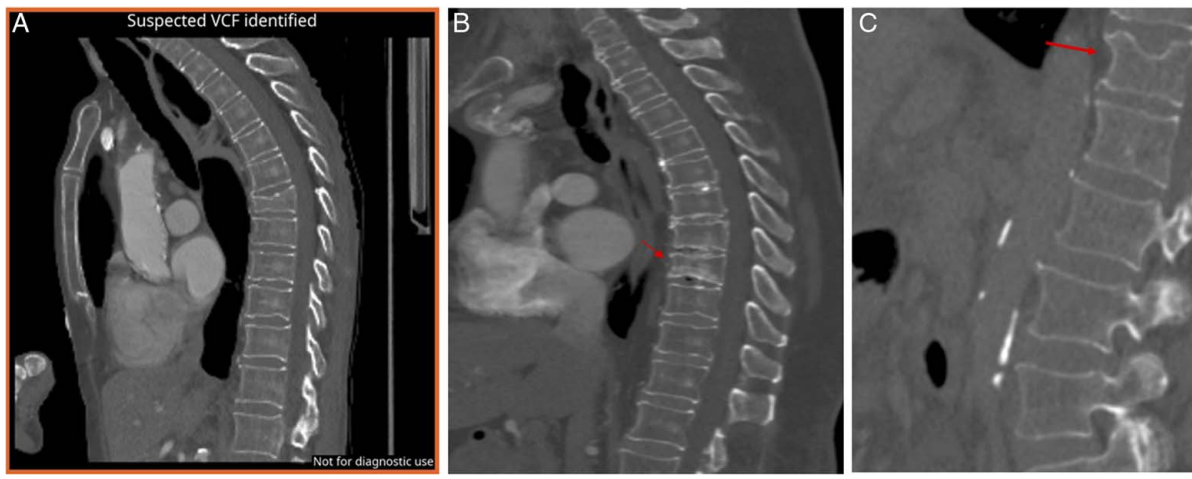

FIGURE 1. Example of a (A) true positive VCF, (B) false negative from mild deformation, and (C) false positive from a Schmorl's node. VCF indicates vertebral compression fracture.

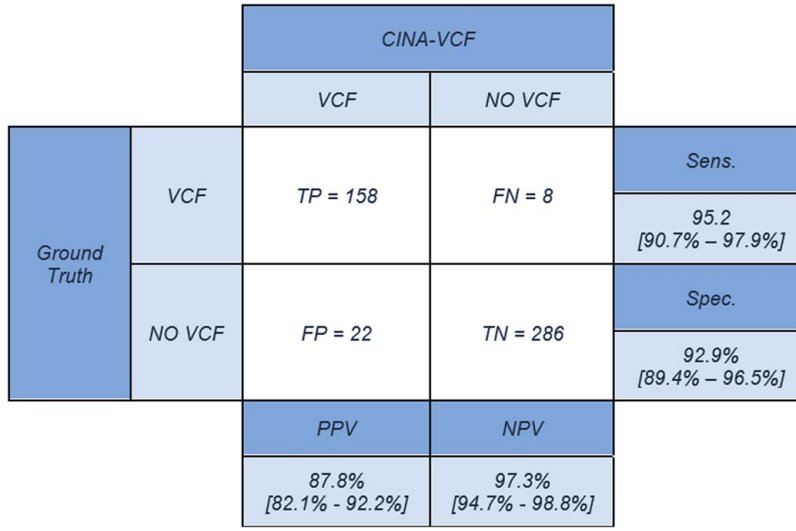

FIGURE 3. Confusion matrix for overall VCF performance. FN indicates false negative; FP, false positive; NPV, negative predictive value; PPV, positive predictive value; Sens, sensitivity; Spec, specificity; TN, true negative; TP, true positive; VCF, vertebral compression fracture.

parameters and patient demographics. Furthermore, CINA-VCF outperformed VCF detection based on clinical reports.

In recent years, AI tools have been applied in evaluating various aspects of VCFs. Prior studies have investigated the use of a machine learning-based VCF algorithm to diagnose moderate to severe grade VCFs on CT, demonstrating sensitivity and specificity of 78% and 87%, respectively, in one study and 73.8% and 92.7% in another study.[17,18] Deep learning algorithms have been shown to be comparable to radiologists in evaluating VCFs.[10] One study utilized a weakly supervised deep learning approach to evaluate multiple VCFs in the lumbar spine on CT, demonstrating equivalent or better performance to a supervised approach.[19] Traditional machine learning techniques combined with bone densitometry have shown high sensitivity in detecting VCFs on CT.[20] Distinguishing acute versus chronic VCF on CT has also been done using a combined machine learning and radiomics approach.[21] Others have used deep learning techniques to detect VCF on radiographs, which are more widely available and lower cost compared with CT or MRI.[22,23]

Strengths of this study include the robust performance metrics, outperforming prior commercially available AI-based VCF detection tools for CT in overall sensitivity and specificity.[17,18] Ground truth was established by neuroradiologists blinded to each other, and CINA-VCF was also blinded to the radiologist's classifications during training and testing. Studies were obtained from multiple scanners and sites, demonstrating the generalizability of the tool in a variety of settings. In addition, CINA-VCF had better performance in detecting VCFs compared with radiologist reports, detecting 88.2% of cases that were initially missed or not reported by the clinical report. This corroborates the increased detection rate from other AI algorithms compared with the radiologist report and shows a potential role for AI in the clinical workspace for aiding the radiologist in incidental VCF detection, which could allow for appropriate follow-up and interventions.[18]

Limitations of the study included the retrospective analysis. VCF mimics and other deformities, such as Schmorl's nodes, intervertebral disc vacuum phenomenon, osteopenia, osteophytes, disc calcifications, ankylosing

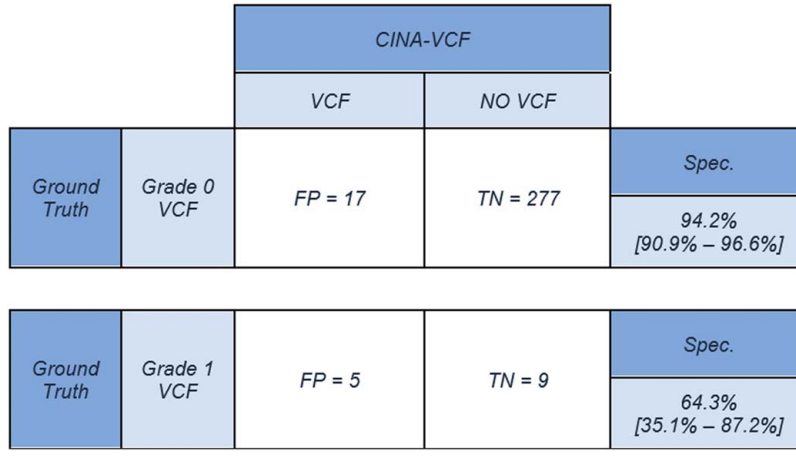

FIGURE 4. Confusion matrices for Genant grade 0 and 1 negative cases.

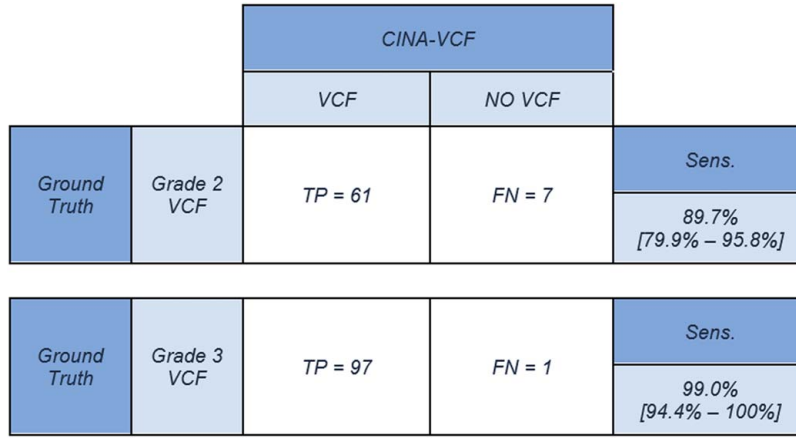

FIGURE 5. Confusion matrices for Genant grade 2 and 3 positive cases.

spondylosis, hemangiomas, and severe scoliosis generated false positives, similar to prior studies.[17] Borderline cases where height loss was near the threshold for positivity was also a limitation. Mild compression deformities were considered negative in this study, even though they are technicality VCFs. The decision to consider Genant grade 1 VCFs as negative cases was made in light of uncertainty concerning misinterpretation of normal variations as vertebral height loss, and the ambiguity of borderline VCF cases likely contributed to the lower specificity of this group. Clinically, management of grades 2 and 3 VCFs is more urgent than grade 1 VCFs, and similar dedicated analyses of moderate to severe VCFs have been performed in prior studies.[16–18] The performance of our method might have been affected by the unequal distribution of our validation data, particularly with respect to VCF grades; however, our goal was to have an unbalanced validation data set, as we aimed to represent the real-world prevalence of VCFs.

Future directions could incorporate other parameters of VCFs to provide a more comprehensive risk profile for current and future osteoporotic fractures. Markers of bone marrow density on CT, including L1 trabecular attenuation values, have been shown to be significantly lower in patients with moderate or severe VCFs and could be added into the algorithm to improve accuracy.[24] Our algorithm graded only the most severe VHL for cases that potentially had multiple VCFs, and future algorithms identifying all VCFs in a single case could be beneficial for counseling patients on future VCF risk and for optimizing therapies. Additional studies should be conducted to evaluate the clinical impact and cost-effectiveness of a VCF tool.

This study validates the applicability and generalizability of an AI-based VCF tool in the clinical setting. CINA-VCF was able to detect VCF with high overall accuracy, even outperforming radiology clinical reports. Implementation of CINA-VCF may help with the identification of patients with osteoporosis and increased fracture risk in patients with moderate to severe VCFs.

REFERENCES

1. Dong Y, Peng R, Kang H, et al. Global incidence, prevalence, and disability of vertebral fractures: a systematic analysis of the global burden of disease study 2019. *Spine J*. 2022;22:857–868.
2. Alexandru D, So W. Evaluation and management of vertebral compression fractures. *Perm J*. 2012;16:46–51.
3. Woo EK, Mansoubi H, Alyas F. Incidental vertebral fractures on multidetector CT images of the chest: prevalence and recognition. *Clin Radiol*. 2008;63:160–164.
4. Williams AL, Al-Busaidi A, Sparrow PJ, et al. Under-reporting of osteoporotic vertebral fractures on computed tomography. *Eur J Radiol*. 2009;69:179–183.
5. Bartalena T, Giannelli G, Rinaldi MF, et al. Prevalence of thoracolumbar vertebral fractures on multidetector CT: under-reporting by radiologists. *Eur J Radiol*. 2009;69:555–559.
6. Urrutia J, Besa P, Piza C. Incidental identification of vertebral compression fractures in patients over 60 years old using computed tomography scans showing the entire thoraco-lumbar spine. *Arch Orthop Trauma Surg*. 2019;139:1497–1503.
7. Löffler MT, Kallweit M, Niederreiter E, et al. Epidemiology and reporting of osteoporotic vertebral fractures in patients with long-term hospital records based on routine clinical CT imaging. *Osteoporos Int*. 2022;33:685–694.
8. Delmas PD, van de Langerijt L, Watts NB, et al. Under-diagnosis of vertebral fractures is a worldwide problem: The IMPACT Study. *J Bone Miner Res*. 2009;20:557–563.
9. Iyer S, Blair A, White C, et al. Vertebral compression fracture detection using imitation learning, patch based convolutional neural networks and majority voting. *Inform Med Unlocked*. 2023;38:101238.
10. Tomita N, Cheung YY, Hassanpour S. Deep neural networks for automatic detection of osteoporotic vertebral fractures on CT scans. *Comput Biol Med*. 2018;98:8–15.
11. Bar A, Wolf L, Amitai OB, et al. Compression fractures detection on CT. *ArXiv*. 2017. https://doi.org/10.48550/arXiv.1706.01671
12. Bendtsen MG, Hitz MF. Opportunistic identification of vertebral compression fractures on CT scans of the chest and abdomen, using an AI algorithm, in a real-life setting. *Calcif Tissue Int*. 2024;114:468–479.
13. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. *ArXiv*. 2015. doi: 10.48550/arXiv.1505.04597
14. Genant HK, Wu CY, van Kuijk C, et al. Vertebral fracture assessment using a semiquantitative technique. *J Bone Miner Res*. 1993;8:1137–1148.
15. 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 (Oxford)*. 2022;61:3269–3278.
16. 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;35:2307–2312.
17. Page JH, Moser FG, Maya MM, et al. Opportunistic CT screening-machine learning algorithm identifies majority of

vertebral compression fractures: a cohort study. *JBMR Plus*. 2023;7:e10778.
18. Pereira RFB, Helito PVP, Leão RV, et al. Accuracy of an artificial intelligence algorithm for detecting moderate-to-severe vertebral compression fractures on abdominal and thoracic computed tomography scans. *Radiol Bras*. 2024;57:e20230102.
19. Choi E, Park D, Son G, et al. Weakly supervised deep learning for diagnosis of multiple vertebral compression fractures in CT. *Eur Radiol*. 2024;34:3750–3760.
20. Burns JE, Yao J, Summers RM. Vertebral body compression fractures and bone density: automated detection and classification on CT images. *Radiology*. 2017;284:788–797.
21. Zhang J, Liu J, Liang Z, et al. Differentiation of acute and chronic vertebral compression fractures using conventional CT based on deep transfer learning features and hand-crafted radiomics features. *BMC Musculoskelet Disord*. 2023;24:165.
22. Paik S, Park J, Hong JY, et al. Deep learning application of vertebral compression fracture detection using mask R-CNN. *Sci Rep*. 2024;14:16308.
23. Xu F, Xiong Y, Ye G, et al. Deep learning-based artificial intelligence model for classification of vertebral compression fractures: a multicenter diagnostic study. *Front Endocrinol (Lausanne)*. 2023;14:1025749.
24. Graffy PM, Lee SJ, Ziemlewicz TJ, et al. Prevalence of vertebral compression fractures on routine CT scans according to L1 trabecular attenuation: determining relevant thresholds for opportunistic osteoporosis screening. *Am J Roentgenol*. 2017;209:491–496.