Improving Radiographic Fracture Recognition Performance and Efficiency Using Artificial Intelligence

Ali Guermazi, MD, PhD • Chadi Tannoury, MD • Andrew J. Kompel, MD • Akira M. Murakami, MD • Alexis Ducarouge, MSc • André Gillibert, MD, MSc • Xinning Li, MD • Antoine Tournier, MSc • Youmna Lahoud, MD • Mohamed Jarraya, MD • Elise Lacave, MSc • Hamza Rahimi, MD • Aloïs Pourchot, MSc • Robert L. Parisien, MD • Alexander C. Merritt, MD • Douglas Comeau, DO • Nor-Eddine Regnard, MD, MSc • Daichi Hayashi, MD, PhD

From the Departments of Radiology (A. Guermazi, A.J.K., A.M.M., H.R., A.C.M., D.H.), Orthopaedic Surgery (C.T., X.L.), and Family Medicine (D.C.), Boston University School of Medicine, Boston, Mass; Department of Radiology, VA Boston Healthcare System, 1400 VFW Parkway, Suite 1B105, West Roxbury, MA 02132 (A. Guermazi); Gleamer, Paris, France (A.D., A.T., E.L., A.P., N.E.); Department of Biostatistics, CHU Rouen, Rouen, France (A. Gillibert); Department of Rheumatology, Harvard Vanguard Medical Associates, Braintree, Mass (Y.L.); Department of Radiology, Musculoskeletal Division, Massachusetts General Hospital, Harvard Medical School, Boston, Mass 02114 (M.J.); Sorbonne Université, CNRS, Institut des Systèmes Intelligents et de Robotique, Paris, France (A.P.); Department of Orthopaedic Surgery, The Mount Sinai Hospital, New York, NY (R.L.P.); University Health Services and Primary Care Sports Medicine, Boston College, Chestnut Hill, Mass (D.C.); and Department of Radiology, Stony Brook University Renaissance School of Medicine, Stony Brook, NY (D.H.). Received April 15, 2021; revision requested June 21; revision received October 13; accepted October 28. **Address correspondence to** A. Guermazi. (e-mail: *guermazi@bu.edu*).

Supported by Gleamer, which developed the AI and software and built the study sample and design.

See also the editorial by Link and Pedoia in this issue.

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

Radiology 2021; 000:1–10 • https://doi.org/10.1148/radiol.210937 • Content codes: MK AI

***Background*:** Missed fractures are a common cause of diagnostic discrepancy between initial radiographic interpretation and the final read by board-certified radiologists.

Purpose: To assess the effect of assistance by artificial intelligence (AI) on diagnostic performances of physicians for fractures on radiographs.

Materials and Methods: This retrospective diagnostic study used the multi-reader, multi-case methodology based on an external multicenter data set of 480 examinations with at least 60 examinations per body region (foot and ankle, knee and leg, hip and pelvis, hand and wrist, elbow and arm, shoulder and clavicle, rib cage, and thoracolumbar spine) between July 2020 and January 2021. Fracture prevalence was set at 50%. The ground truth was determined by two musculoskeletal radiologists, with discrepancies solved by a third. Twenty-four readers (radiologists, orthopedists, emergency physicians, physician assistants, rheumatologists, family physicians) were presented the whole validation data set (n = 480), with and without AI assistance, with a 1-month minimum washout period. The primary analysis had to demonstrate superiority of sensitivity per patient and the noninferiority of specificity per patient at −3% margin with AI aid. Stand-alone AI performance was also assessed using receiver operating characteristic curves.

Results: A total of 480 patients were included (mean age, 59 years ± 16 [standard deviation]; 327 women). The sensitivity per patient was 10.4% higher (95% CI: 6.9, 13.9; $P < .001$ for superiority) with AI aid (4331 of 5760 readings, 75.2%) than without AI (3732 of 5760 readings, 64.8%). The specificity per patient with AI aid (5504 of 5760 readings, 95.6%) was noninferior to that without AI aid (5217 of 5760 readings, 90.6%), with a difference of +5.0% (95% CI: +2.0, +8.0; P = .001 for noninferiority). AI shortened the average reading time by 6.3 seconds per examination (95% CI: −12.5, −0.1; P = .046). The sensitivity by patient gain was significant in all regions (+8.0% to +16.2%; $P < .05$) but shoulder and clavicle and spine (+4.2% and +2.6%; P = .12 and .52).

Conclusion: AI assistance improved the sensitivity and may even improve the specificity of fracture detection by radiologists and nonradiologists, without lengthening reading time.

Online supplemental material is available for this article.

Fracture detection using radiography is one of the most common tasks in patients with high- or low-energy trauma in various clinical settings, including the emergency department, urgent care, and outpatient clinics such as orthopedics, rheumatology, and family medicine. Missed fractures on radiographs are one of the most common causes of diagnostic discrepancies between initial interpretations by nonradiologists or radiology residents and the final read by board-certified radiologists, leading to preventable harm or delay in care to the patient (1–3). Fracture interpretation errors can represent up to 24% of harmful diagnostic errors seen in the emergency department (2). Furthermore, inconsistencies in radiographic diagnosis of fractures are more common during the evening and overnight hours (5 PM to 3 AM), likely related to nonexpert reading and fatigue (3). In patients with multiple traumas, the proportion of missed injuries, including fractures, can be high on the forearm and hands (6.6%) and feet (6.5%) (4,5).

To date, several studies about artificial intelligence (AI) aid to fracture detection have been performed focusing only on certain body parts, such as hand, wrist, and forearm (6–9); hip and pelvis (10,11); knees (9); and spine (12). One study evaluated fractures in 11 body locations,

Abbreviations

AI = artificial intelligence, AUC = area under the receiving operating characteristic curve, ROC = receiver operating characteristic

Summary

Artificial intelligence assistance for searching skeletal fractures on radiographs improved the sensitivity and specificity of readers and shortened their reading time.

Key Results

- In a retrospective study of 480 patients, artificial intelligence (AI)–assisted radiographic interpretation by six types of readers showed a 10.4% improvement of fracture detection sensitivity (75.2% vs 64.8%, superiority $P < .001$) without specificity reduction (5.0%, 95.6% vs 90.6%; $P = .001$ for noninferiority).
- AI assistance shortened the radiograph reading time by 6.3 seconds per patient ($P = .046$).
- The improvement in sensitivity was significant in all locations (delta mean, 8.0%–16.2%; $P < .05$) but shoulder and clavicle and thoracolumbar spine (delta mean, 4.2% and 2.6%; $P = .12$ and .52, respectively).
- The stand-alone performance of the AI algorithm for fracture detection had an area under the receiver operating characteristic curve of 0.97 ($P < .001$).

including the upper and lower extremities and spine (13), but the clinicians who read the radiographs with AI and without AI assistance were emergency medicine physicians and physician assistants only, with senior orthopedic surgeons providing the ground truth; no radiologist was involved in the radiographic interpretation. Another recent study analyzed fractures in 16 anatomic locations; however, readers of the radiographs were radiologists and orthopedic surgeons only (14).

The aim of this study was to assess the effect of assistance by AI on diagnostic performances of physicians for fractures on radiographs.

Materials and Methods

This retrospective study was funded by Gleamer, which developed the AI and software and built the study sample and design. Data analysis and manuscript writing were performed by authors not affiliated with Gleamer (A. Gillibert, a biostatistician, and A. Guermazi and D.H., musculoskeletal radiologists). Five authors (A.D., A.T., E.L., A.P., and N.E.R.) are employees of Gleamer. Two authors (A. Guermazi and D.H.) had control of the data and the information submitted for publication. The current protocol was approved by the WellCare Group institutional review board (no. 20202256), who waived informed consent because of the retrospective nature of this study and the fact that all images were deidentified and without any clinical information. Our study was Health Insurance Portability and Accountability Act compliant.

Study Design

The AI algorithm was developed using a development data set of 60 170 radiographs of patients with trauma from 22 institutions between January 2011 and May 2019. This data set was randomly split into a training set (70%), validation set (10%), and internal test set (20%). A deep learning algorithm was trained and validated on this development data set based on the "Detectron2" framework (15), which was further revised and tailor made to the analysis of radiographs by Gleamer. Notably, data augmentation—by random rotation, flipping, translation, cropping, and resizing—was performed during training. We used 270 000 iterations to train the model ("training batch") and updated the parameters using a stochastic gradient algorithm with a batch size of four. Depending on the threshold set on the scores returned by the algorithm for each region of interest, the final pipeline is capable of running at different operating points in terms of sensitivity and specificity. When the AI confidence level surpasses the threshold, the software aid highlights the region of interest with a white square box on the radiograph.

The AI was validated for detection and localization of fractures on digital radiographs of diagnostic quality and subsequently integrated into a radiologic image interpretation software developed by Gleamer as a tool to assist fracture detection, highlighting potential fractures with a rectangular box.

The current study is a retrospective diagnostic study using the multi-reader, multi-case methodology, based on an external multicenter data set from the United States, registered at ClinicalTrials.gov (identification no: NCT04532580). The overall study design is summarized in Figure 1.

The code underlying this work can be found online at *https://github.com/facebookresearch/detectron2*.

Data Set Acquisition

A total of 480 radiographic examinations were acquired through three radiologic data providers in the United States. Our data set was generated from 11 different manufacturers of radiologic data sources: Konica Minolta, Samsung Electronics, GE Healthcare, Philips Medical Systems, KODAK, Canon, Swissray, Hologic, Varian, Siemens, and Fujifilm. Images were acquired using these instruments and collected from multiple institutions within the United States. Between July 2020 and January 2021, the radiographic examinations were sampled following inclusion and exclusion criteria with stratification on region and fracture status. At least 60 examinations had to be included for each of the following anatomic regions: foot and ankle, knee and leg, hip and pelvis, hand and wrist, elbow and arm, shoulder and clavicle, rib cage, and thoracolumbar spine. In addition, 50% of examinations of each region with one fracture or more and 50% of examinations with no fracture had to be included, as well as 25% of examinations with at least one "nonobvious" fracture, according to experts during ground truth determination, and 25% with only "obvious" fractures. The number of projections or hanging protocol was variable because images were retrospectively collected from multiple institutions with their own image acquisition protocols outside the control of the investigators.

Ground Truth Definition

Ground truth was established by two experienced musculoskeletal radiologists (D.H. and A.J.K., with 12 years and 8 years of experience, respectively), who independently interpreted all examinations without clinical information. Only acute fractures were considered to be a positive finding in our study. Therefore, the term *fracture* refers to an acute fracture, unless otherwise stated.

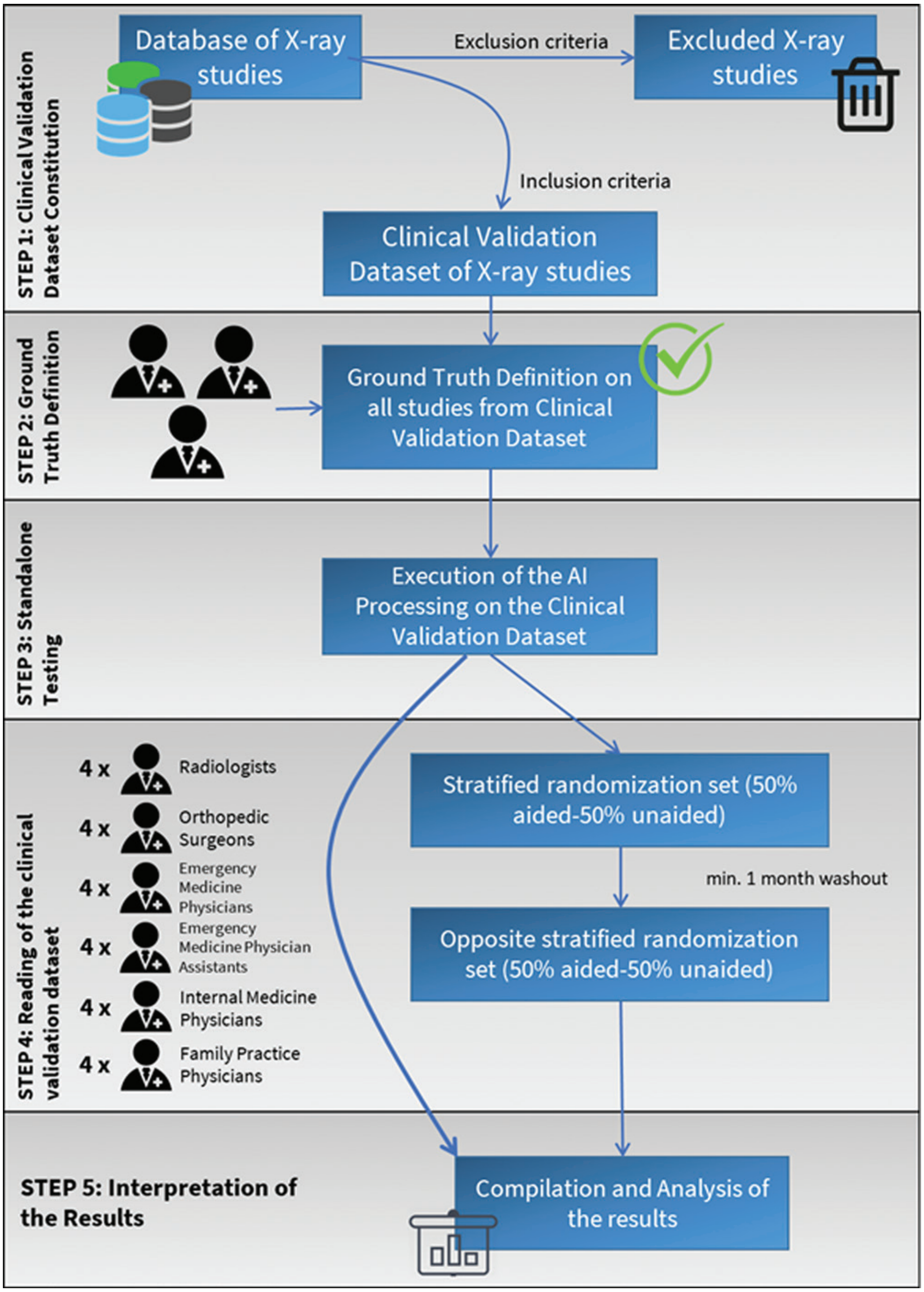
STEP 1: Clinical Validation Dataset Constitution
Database of X-ray studies
Exclusion criteria
Excluded X-ray studies
Inclusion criteria
Clinical Validation Dataset of X-ray studies
STEP 2: Ground Truth Definition
Ground Truth Definition on all studies from Clinical Validation Dataset
STEP 3: Standalone Testing
Execution of the AI Processing on the Clinical Validation Dataset
STEP 4: Reading of the clinical validation dataset
4 x Radiologists
4 x Orthopedic Surgeons
4 x Emergency Medicine Physicians
4 x Emergency Medicine Physician Assistants
4 x Internal Medicine Physicians
4 x Family Practice Physicians
Stratified randomization set (50% aided-50% unaided)
min. 1 month washout
Opposite stratified randomization set (50% aided-50% unaided)
STEP 5: Interpretation of the Results
Compilation and Analysis of the results

Figure 1: Flowchart summarizes the study design. min. = minimum.

Each acute fracture was tagged as "obvious" or "nonobvious" by each expert. Examinations with at least one nonobvious fracture were considered as nonobvious. "Obvious fractures" were fractures that were displaced, comminuted, markedly angulated, or otherwise easily identifiable at a glance on radiographs for any reader. "Nonobvious fractures" were fractures that were subtle nondisplaced fractures that required careful attention and detailed analysis of the radiograph even by expert radiologists.

Clinical Validation Execution

All examinations were independently interpreted by 24 clinicians from multiple institutions within the United States, including both in-training and board-certified physicians with variable years of experience (2–18 years) in radiographic interpretation for fracture detection (see Table E1 [online] for details).

All readers were presented, in random order, the 480 radiographic examinations of the validation data set twice—once with the assistance of AI software and once without the assistance, with a minimum washout period of 1 month.

Reader Performance

The sensitivity per patient was defined as the proportion of patients for whom all actual fractures were discovered (each one, on at least one radiographic view), including potentially multiple fractures at more than one region, among patients having at least one fracture, even if some incorrect marks (false positives) had been added by the reader. The specificity per patient was defined as the proportion of patients for whom no fracture mark was placed by the reader among patients having no fracture.

The sensitivity per fracture was defined as the proportion of fractures correctly detected by the reader among all fractures, counting multiple fractures per patient where appropriate. The per-patient average number of false-positive fractures was defined as the average number of marks put outside of a fracture per patient. The Youden index per patient was defined as (sensitivity per patient) + (specificity per patient) − 1.

Statistical Analysis

The two coprimary outcomes were sensitivity per patient and specificity per patient. For the trial to be successful, both the superiority of the sensitivity per patient and the noninferiority of the specificity per patient, at a margin of −3%, had to be demonstrated for AI-aided readings compared with unaided readings, on the set of 24 readers, with one-sided significance thresholds set at 2.5%. Assuming that the clinical consequences of a false-negative finding are twice as bad as the consequences of a false-positive finding, a prevalence of fractures of 20%, and a sensitivity gain of 8.7% (observed in a previously published study [16]), a specificity inferiority of 4.35% would be acceptable; a noninferiority margin at −3% is safer, taking into account the subjectivity of the 2:1 false-negative–to–false-positive ratio.

In case of successful analysis, the superiority on specificity per patient was searched (hierarchical testing). The effect of AI aid on sensitivity per patient and specificity per patient and other metrics was assessed with a modified paired Student *t* test taking in account the correlation between readers and between examinations and interactions (Appendix E1 [online]). Sensitivity analyses are described in Appendix E1 (online). A similar method was used to assess the period and the carryover effects. In addition to evaluation of diagnostic performance of the human readers with or without AI assistance, stand-alone AI performance was also assessed using receiver operating characteristic (ROC) curves and free-response ROC curves (Appendix E1 [online]).

Statistical analyses were all performed using software (R, version 4.0.2; the R Foundation for Statistical Computing). The significance threshold was set at two-sided 5% ($P < .05$) for all secondary analyses, without multiple testing procedure.

Results

Data Set Characteristics

A total of 480 patients were included (Fig 2) (mean age, 59 years ± 16 [standard deviation]; 327 women) (Table 1). Becausee some patients had several fractures, a total of 350 fractures were found in 240 patients. Precise anatomic locations of all fractures are summarized in Table E2 (online).

Primary Analysis

The sensitivity per patient was estimated at 64.8% (3732 of 5760 readings) without AI aid and 75.2% (4331 of 5760 readings) with AI aid, with an AI effect estimated at +10.4% (95% CI: 6.9, 13.9; one-sided $P < .001$ for superiority). The specificity per patient was estimated at 90.6% (5217 of 5760 readings) without AI aid and 95.6% (5504 of 5760 readings) with AI aid, with an AI effect estimated at +5.0% (95% CI: +2.0, +8.0; one-sided $P = .001$ for noninferiority and one-sided $P = .001$ for superiority). Therefore, the primary analysis was successful, and the superiority of specificity per patient with AI aid was demonstrated (hierarchical test). Other metrics for all readers are compared in Table 2 by using two-sided tests. Overall, the reading time was 6.3 seconds shorter (95% CI: −12.5, −0.1; $P = .046$) with AI aid than without.

Sensitivity Analyses

Planned sensitivity analyses did not change the point estimate and tended to reduce the width of the CI of the sensitivity per patient gain, up to 26% for the linear mixed-effects model compared with the primary analysis, meaning that the primary analysis is more conservative (Table E3 [online]). A post hoc sensitivity analysis removing the outlier (a rheumatologist with extremely low diagnostic performances) did not change the conclusions: There was a 9.5% (95% CI: +6.4, +12.5; one-sided $P < .001$ for superiority) gain of sensitivity per patient and 4.0% (95% CI: +1.8, +6.2; one-sided $P < .001$ for superiority) gain of specificity per patient.

Period and Carryover Effects

The carryover effect (time × AI interaction) for the sensitivity per patient was estimated as an increase of effect of +2.41% (95% CI: −7.10, +2.28; $P = .30$) when the AI assistance was in the second period rather than in the first period. The period effect, equal to the average sensitivity per

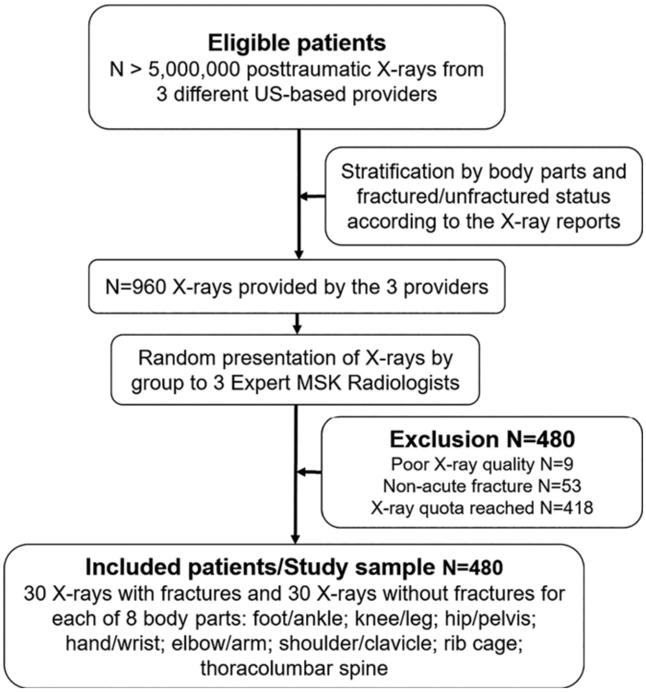
Eligible patients
N > 5,000,000 posttraumatic X-rays from 3 different US-based providers
Stratification by body parts and fractured/unfractured status according to the X-ray reports
N=960 X-rays provided by the 3 providers
Random presentation of X-rays by group to 3 Expert MSK Radiologists
Exclusion N=480
Poor X-ray quality N=9
Non-acute fracture N=53
X-ray quota reached N=418
Included patients/Study sample N=480
30 X-rays with fractures and 30 X-rays without fractures for each of 8 body parts: foot/ankle; knee/leg; hip/pelvis; hand/wrist; elbow/arm; shoulder/clavicle; rib cage; thoracolumbar spine

Figure 2: Flowchart of the study sample determination. MSK = musculoskeletal.

Table 1: Demographics by Anatomic Location

Anatomic Location	Positive for Fracture		Negative for Fracture		Total	
	No. of Women	Age (y)*	No. of Women	Age (y)*	No. of Women	Age (y)*
Foot and ankle	20/30 (67)	49 ± 14	22/30 (73)	60 ± 15	42/60 (70)	54 ± 16
Knee and leg	17/27 (63)	59 ± 13	17/30 (57)	64 ± 14	34/57 (60)	62 ± 14
Hip and pelvis	19/30 (63)	56 ± 12	25/30 (83)	65 ± 16	44/60 (73)	61 ± 15
Hand and wrist	22/30 (73)	51 ± 17	19/30 (63)	62 ± 16	41/60 (68)	56 ± 17
Elbow and arm	19/30 (63)	50 ± 13	24/28 (86)	63 ± 14	43/58 (74)	56 ± 15
Shoulder and clavicle	19/29 (66)	63 ± 19	13/25 (52)	58 ± 17	32/54 (59)	61 ± 18
Rib cage	22/29 (76)	67 ± 12	19/30 (63)	67 ± 15	41/59 (70)	67 ± 14
Thoracolumbar spine	21/30 (70)	46 ± 16	18/26 (69)	68 ± 16	39/56 (70)	56 ± 20
Multiple locations	4/5 (80)	60 ± 27	7/11 (64)	67 ± 10	11/16 (69)	65 ± 17
All regions	163/240 (68)	55 ± 16	164/240 (68)	64 ± 15	327/480 (68)	59 ± 16

Note.—Numbers in parentheses are percentages.

* Data are means ± standard deviations.

Table 2: Diagnostic Performance of 24 Readers for Fracture Detection with and without AI Assistance

Parameter	Without AI (*n* = 480)*	With AI (*n* = 480)*	Delta Mean†
Sensitivity per patient	64.8 ± 9.4	75.2 ± 5.9	10.4 (6.9, 13.9) [< .001]
Specificity per patient	90.6 ± 8.0	95.6 ± 2.8	5.0 (2.0, 8.0) [.001]
Youden index per patient	55.4 ± 15.0	70.7 ± 6.0	15.4 (15.4, 15.4) [< .001]
Average fracture detection sensitivity per patient	65.4 ± 10.6	76.2 ± 6.4	10.8 (7.1, 14.4) [< .001]
Average no. of false-positive fractures per patient in patients without fracture	0.11 ± 0.10	0.05 ± 0.04	−0.06 (−0.10, −0.02) [.002]
Average no. of false-positive fractures per patient in patients with fracture	0.10 ± 0.06	0.10 ± 0.06	0.003 (−0.026, 0.031) [.85]
Reading time (sec)	55.5 ± 32.6	49.2 ± 28.5	−6.3 (−12.5, −0.1) [.046]

Note.—Except where indicated, data are percentages. AI = artificial intelligence.

* Data are means ± standard deviations.

† Numbers in parentheses are 95% CIs. Numbers in brackets are *P* values.

patient in the second period minus the average sensitivity per patient in the first period, was absent (−0.19%; 95% CI: −2.03, +1.65; *P* = .83). There was no evidence of carryover effect for the reading time (time × AI interaction: +3.0 seconds [95% CI: −2.6, +8.6; *P* = .28]), but the period effect was present (−7.9 [95% CI: −12.6, −3.2; *P* = .002]) without generating a bias in our main analysis due to the balanced design of the study.

Subgroup Analyses

Results of readings by all 24 readers are presented in Table E2 (online). As shown in Table 3, overall, AI-assisted reading resulted in improved sensitivity per patient for all specialties (ranging from +9.1% to +17.2%, *P* values ranging from .02 to .046) but radiologists and rheumatologists (+7.6% [95% CI: −3.8, 19.0; *P* = .12] and +17.2% [95% CI: −0.4, 34.8; *P* = .05], respectively). Although the specificity per patient was higher with AI aid in all specialties, we found no evidence of differences (ranging from +2.0% to +14.0%, *P* values ranging from .07 to .49); the statistical precision was low for these analyses. We found no evidence that the effect of the AI assistance was different between specialties for sensitivity per patient (*P* = .39 for interaction) or for specificity per patient (*P* = .14 for interaction).

The AI aid improved the sensitivity per patient from 58.0% (1975 of 3408 readings) to 70.4% (2398 of 3408 readings) for nonobvious fractures (+12.4%; 95% CI: 7.8, 17.0; *P* < .001) and from 74.7% (1757 of 2352 readings) to 82.2% (1933 of 2352 readings) for obvious fractures (+7.5%; 95% CI: 3.7, 11.2; *P* < .001), with no evidence of a difference of gain (−4.9%; 95% CI: −9.9, +0.0; *P* = .05).

Overall, AI-aided reading showed improvement of sensitivity per patient for fracture detection across all anatomic locations (with greater than 10% improvement observed in sensitivity per patient for fractures in foot and ankle, knee and leg, hip and pelvis, hand and wrist, elbow and arm, and rib cage) but thoracolumbar spine and shoulder and clavicle (less than 5% improvement, no evidence of difference), as detailed in Table E4 (online), and for both expert and nonexpert readers (Table 4). We found neither evidence of an interaction of the effect of AI assistance on the sensitivity per patient (between-location standard deviation of AI effect = 4.4%; *P* = .30) with the location, nor for the specificity per patient (between-location standard deviation of AI effect = 1.7%,

Table 3: Improvement in Diagnostic Performance of Fracture Detection with AI Assistance

Reader	Absolute Gain of Sensitivity per Patient (Sensitivity) (%)	Absolute Gain of Specificity per Patient (Specificity) (%)	Relative Change in Average No. of False-Positive Fractures per Patient in Patients with Fracture (%)	Relative Change in Average No. of False-Positive Fractures per Patient in Patients without Fracture (%)
Four radiologists	7.6 (−3.8, 19.0) [.12]	2.8 (−8.6, 14.2) [.49]	−16 (−61, 81) [.52]	−45 (−97, 1044) [.58]
Four orthopedists	9.1 (0.7, 17.4) [.04]	2.0 (−3.0, 7.0) [.29]	18 (−34, 111) [.44]	−26 (−62, 46) [.26]
Four emergency physicians	9.9 (3.2, 16.6) [.02]	3.4 (−2.0, 8.8) [.14]	20 (−15, 69) [.19]	−59 (−97, 394) [.34]
Four emergency medicine physician assistants	9.4 (0.3, 18.4) [.046]	2.5 (−0.5, 5.5) [.08]	45 (−66, 531) [.48]	−50 (−83, 44) [.13]
Four rheumatologists	17.2 (−0.4, 34.8) [.05]	14.0 (−3.9, 31.9) [.09]	−18 (−70, 122) [.57]	−64 (−97, 337) [.29]
Four family physicians	9.3 (3.4, 15.2) [.02]	5.2 (−0.9, 11.3) [.07]	7 (−52, 137) [.80]	−60 (−93, 146) [.21]
All readers	10.4 (6.9, 13.9) [< .001]	5.0 (2.0, 8.0) [.002]	3 (−23, 37) [.85]	−53 (−76, −8) [.03]

Note.— Numbers in parentheses are 95% CIs, and numbers in brackets are *P* values. AI = artificial intelligence.

Table 4: Expert Readers versus Nonexpert Readers: Subgroup Analysis

Reader and Parameter	Without AI (*n* = 480)	With AI (*n* = 480)	Delta Mean*
Nonexpert			
Sensitivity per patient (%)	61.8 ± 8.6	73.3 ± 5.2	11.4 (7.0, 15.9) [< .001]
Specificity per patient (%)	89.8 ± 8.7	96.1 ± 2.7	6.3 (2.3, 10.3) [.004]
Youden index per patient (%)	51.6 ± 15.4	69.3 ± 5.4	17.7 (10.6, 24.8) [< .001]
Sensitivity per fracture (%)	62.0 ± 9.6	74.3 ± 5.8	12.3 (7.7, 16.9) [< .001]
Average no. of false-positive fractures per patient in patients without fracture	0.12 ± 0.11	0.05 ± 0.04	−0.07 (−0.12, −0.03) [.005]
Average no. of false-positive fractures per patient in patients with fracture	0.09 ± 0.06	0.09 ± 0.06	0.004 (−0.031, 0.039) [.79]
Time (sec)†	54.1 ± 33.8	49.3 ± 31.7	−4.8 (−12.5, 2.9) [.21]
Expert			
Sensitivity per patient (%)	70.7 ± 8.3	79.1 ± 5.5	8.3 (3.2, 13.5) [.006]
Specificity per patient (%)	92.1 ± 6.5	94.5 ± 2.8	2.4 (−2.0, 6.8) [.23]
Youden index per patient (%)	62.8 ± 11.5	73.5 ± 6.6	10.7 (3.5, 18.0) [.01]
Sensitivity per fracture (%)	72.2 ± 9.4	79.9 ± 6.1	7.7 (2.6, 12.8) [.009]
Average no. of false-positive fractures per patient in patients without fracture	0.09 ± 0.08	0.06 ± 0.04	−0.03 (−0.09, 0.02) [.21]
Average no. of false-positive fractures per patient in patients with fracture	0.19 ± 0.06	0.12 ± 0.05	−0.001 (−0.045, 0.043) [.96]
Time (sec)†	58.5 ± 32.0	49.1 ± 22.6	−9.4 (−22.3, 3.6) [.13]

Note.—Data are for 480 patients. Expert readers are orthopedists and radiologists. Nonexpert readers are all others. Except where indicated, data are means ± standard deviations. AI = artificial intelligence.

* Numbers in parentheses are 95% CIs, and numbers in brackets are *P* values.

† Data are averages ± standard deviations.

P = .77). There were 171 patients with a single fracture and 69 patients with multiple fractures, and a similar degree of sensitivity improvement was observed for both subgroups. Sensitivity per fracture improved from 76.0% (3120 of 4104 readings) to 86.1% (3533 of 4104 readings), that is, by 10.1% (95% CI: 6.2, 13.9; *P* < .001) for patients with single fractures, and from 55.3% (2375 of 4296 readings) to 66.7% (2867 of 4296 readings), that is, by 11.5% (95% CI: 6.7, 16.2; *P* < .001) for patients with multiple fractures. The difference of AI gain of sensitivity per fracture between patients with multiple fractures (+11.5%) and single fractures (+10.1%) was estimated at 1.4% (95% CI: −3.3, –6.0; *P* = .54).

One outlier rheumatologist reader had a very low sensitivity per patient without AI assistance (37.9%), which was the lowest value among all readers, and AI assistance improved sensitivity to 70.0% (gain of 32.1%). When this reader was excluded, the other three rheumatologist readers had an average of 12.2% (95% CI: −1.7, +26.1; *P* = .06) improvement in sensitivity per patient, which is closer to 10.4% improvement of all readers. The outlier rheumatologist had

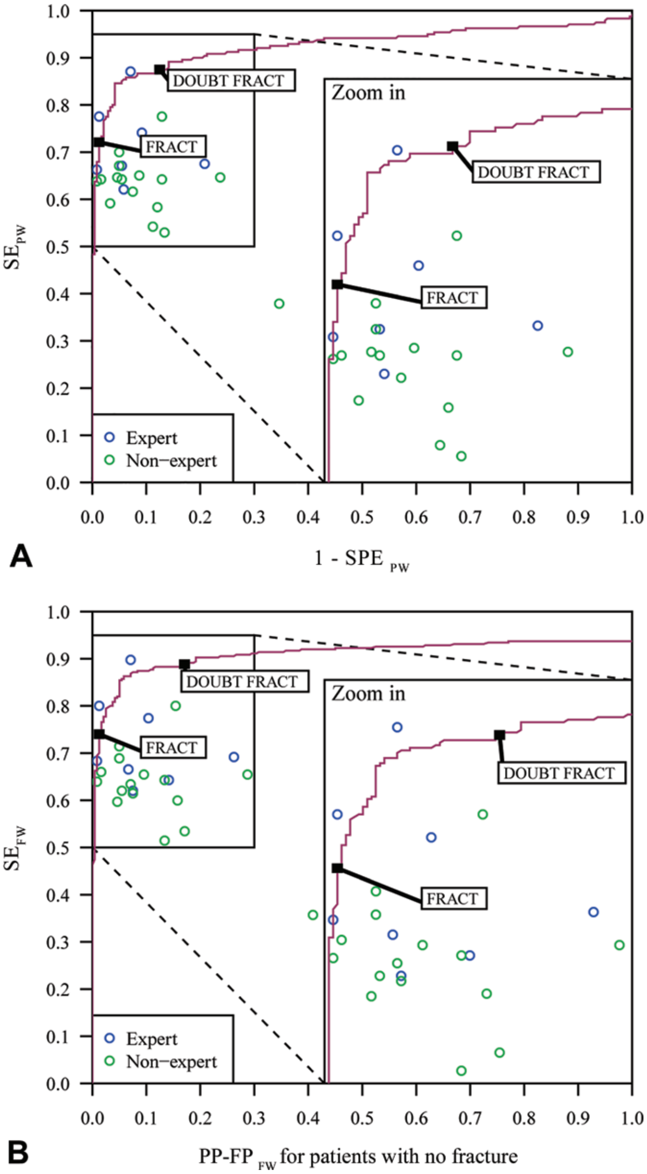

Figure 3: **(A)** Stand-alone artificial intelligence (AI) receiver operating characteristic (ROC) and **(B)** free-response ROC curves. Expert readers (blue circles) include orthopedists and radiologists, and nonexpert readers (green circles) comprise all other readers. AI-unassisted readers' diagnostic performances are shown in both graphs (blue and green circles). Area under the ROC curve for the stand-alone AI (solid line in **A**) was 0.93 (95% CI: 0.90, 0.95). Note: "Free-response ROC curve" is a modification of ROC curve to adapt to multiple fractures in one patient, with per-fracture metrics rather than per-patient metrics. PP-FP$_{FW}$ = average number of false-positive fractures per patient, SE$_{PW}$ = average fracture detection sensitivity per patient, SPE$_{PW}$ = average fracture detection specificity per patient.

a very low specificity per patient without AI (65.4%) and had a 28.3% improvement with AI.

Evaluation of Stand-Alone AI Performance

Stand-alone AI ROC and free-response ROC curves are shown in Figure 3. For stand-alone AI, the area under the ROC curve (AUC) based on custom metrics was 0.93 (95% CI: 0.90, 0.95; $P < .001$) and 0.97 (95% CI: 0.95, 0.98; $P < .001$) for an AUC based on a binary diagnosis of fracture to be comparable to the literature. The performances of the stand-alone AI at the high-sensitivity threshold named DOUBT-FRACT for ribs and thoracolumbar spine were lower than the stand-alone performance for other anatomic locations, with a relatively large number of false-positive findings (Tables 4, 5). There was no evidence that AI assistance provided a gain of sensitivity per patient in the thoracolumbar spine (from 59.5% [371 of 624 readings] to 62.0% [387 of 624 readings], +2.6%, $P = .56$), but it provided a major gain in the rib cage (from 29.2% [210 of 720 readings] to 45.4% [327 of 720 readings], +16.2%; $P = .002$) (Table E4 [online]).

Examples of true- and false-positive and/or negative examinations are shown in Figures 4 and 5. During our study, a newer version of the AI algorithm was developed (although not included in the software), but we found no evidence of difference between the new and old AI algorithm (custom metrics AUC, 0.92; 95% CI: 0.89, 0.94) with a difference estimated at +0.002 (95% CI: −0.006, +0.009; $P = .69$) compared to the original AI algorithm.

Discussion

Missed fractures on radiographic images are not an uncommon problem in the setting of acute trauma, and we aimed to assess the effect of artificial intelligence (AI) assistance on diagnostic performances of physicians for radiographic fracture detection. We used an external multicenter data set from the United States, including multivendor radiographic acquisition systems that were not related to the development set originating from Europe, providing the robust generalization capacity of the model. Our AI system can interpret full-size high-spatial-resolution images, including multiple radiographic views in a patient, and can be integrated into picture archiving and communication systems used in the daily clinical practice.

The stand-alone performance of our AI algorithm (AUC, 0.97) is comparable to that of other published studies (AUC >0.90 for most studies) (6–13). In this retrospective study of 480 patients, AI-assisted radiographic reading by six types of readers showed a 10.4% improvement of fracture detection sensitivity (75.2% vs 64.8%; $P < .001$ for superiority) without specificity reduction (+5.0%; 95.6% vs 90.6%, $P = .001$ for noninferiority). AI assistance shortened the radiograph reading time by 6.3 seconds per patient ($P = .046$). The improvement in sensitivity was significant in all locations (delta mean, 8.0%–16.2%; $P < .05$) but shoulder and clavicle and thoracolumbar spine (delta mean: 4.2% and 2.6%, respectively; $P = .12$ and .52).

A major advantage that AI can bring to clinical practice, particularly in the emergency setting, is its potential to function as a triage system at busy medical centers. If the AI can detect a fracture prior to radiologists' interpretation, then that particular study can become prioritized on the work list. If radiologists can prioritize reading studies with a potentially positive finding, then delay between initial nonexpert reading and the radiologists' final report can be minimized, thereby improving the patient care. Another potential benefit of AI is shorter reading time. Even if by only a few seconds per radiographic examination, a reduction in reading time can add up to a meaningful amount of time saved for radiologists who may read 200–300 radiographs per day. However, we cannot prove this will really be the case in real-life situations. The AI-assisted fracture recognition also has the

Table 5: Stand-Alone AI Performance for Fracture Detection Using DOUBT-FRACT Threshold

Anatomic Location	Sensitivity per Patient (%)*	Specificity per Patient (%)*	Average No. of False-Positive Fractures per Patient in Patients without Fracture†	Average No. of False-Positive Fractures per Patient in Patients with Fracture†	Patient-wise AUC‡
Foot and ankle	93 (28/30)	93 (28/30)	0.07 ± 0.25	0.20 ± 0.55	0.97 (0.94, 0.99)
Knee and leg	90 (27/30)	93 (25/27)	0.07 ± 0.27	0.07 ± 0.25	0.93 (0.85, 0.98)
Hip and pelvis	90 (27/30)	87 (26/30)	0.23 ± 0.68	0.20 ± 0.48	0.93 (0.85, 0.98)
Hand and wrist	93 (28/30)	100 (30/30)	0.00 ± 0.00	0.13 ± 0.57	0.94 (0.83, 0.98)
Elbow and arm	100 (28/28)	97 (29/30)	0.03 ± 0.18	0.04 ± 0.19	0.98 (0.96, 0.99)
Shoulder and clavicle	84 (21/25)	83 (24/29)	0.17 ± 0.38	0.08 ± 0.28	0.90 (0.79, 0.96)
Rib cage	77 (23/30)	69 (20/29)	0.55 ± 0.95	0.50 ± 0.90	0.75 (0.60, 0.87)
Thoracolumbar spine	77 (20/26)	80 (24/30)	0.23 ± 0.50	0.38 ± 0.90	0.86 (0.73, 0.95)
Multiple locations	73 (8/11)	80 (4/5)	0.20 ± 0.45	1.64 ± 3.04	0.75 (0.56, 0.93)
All locations	88 (210/240)	88 (210/240)	0.17 ± 0.51	0.27 ± 0.90	0.93 (0.90, 0.95)

Note.—AI = artificial intelligence, AUC = area under the receiver operating characteristic curve.

* Numbers in parentheses are patients.

† Data are means ± standard deviations.

‡ Numbers in parentheses are 95% CIs.

potential to enhance diagnostic ability of both radiologists and nonradiologists, not only by detecting subtle findings difficult to visualize with human eyes but also by preventing cognitive errors due to human fatigue or satisfaction bias in image interpretation (17,18). The benefit of AI may be especially seen in emergency medicine physicians and physician assistants, on-call orthopedic surgeons, and on-call radiologists likely exposed to system-related errors in radiographic interpretation, such as visual fatigue and decision fatigue (19–21).

AI assistance was helpful for detection of nonobvious or subtle fractures but also for detection of obvious fractures, which was unexpected. This confirms clinical usefulness of AI in real clinical practice. In contrast, sensitivity and specificity for detection of rib and thoracolumbar spine fractures were relatively low for both AI-assisted human readings (sensitivity per patient, 45.4% for ribs and 62.0% for spine) and the stand-alone AI algorithm (sensitivity per patient, 76.7% for ribs and 76.9% for spine). The stand-alone AI outperformed human readers in both instances. Detection of fractures in multiple anatomic locations in a single patient was also a relative weakness of all human readers, as well as the stand-alone AI algorithm.

Ongoing research has shown that an AI algorithm can be used to provide a percentage likelihood or risk score of a specific pathologic condition, such as likelihood of cancer, based on imaging data as well as other clinical information (22). In our study, the AI provided a confidence level but not a precise probability of fracture. The probability of fracture is highly dependent on the clinical setting, and, thus, it may be hard to provide correct figures.

Our study had limitations. First, it was retrospective in nature and all radiographs were read without relevant clinical information (23). In a real clinical setting, nonradiologist clinicians can examine the patient and obtain detailed history to identify the area of concern before looking at the radiographs, thus improving both diagnostic sensitivity and specificity. Conversely, radiologists often interpret radiographs with little or inadequate clinical history in real life, thus naturally performing a somewhat "blinded" interpretation. As far as radiologists are concerned, therefore, blinded reading in this study may not necessarily be completely unrealistic. During the radiographic interpretation by human readers, there were several instances where it was difficult to determine whether the visible fracture was acute or chronic due to lack of clinical information, particularly involving ribs and thoracolumbar spine. A contextual bias could alter the interpretation of readers because the setting was quite different from real life, especially for nonradiologist readers (24). Second, due to artificially set 50% prevalence of fractures in our study sample, it was not possible to calculate positive or negative predictive values. The artificial balance between anatomic locations and reader specialties made the sample nonrepresentative of the actual population who will benefit from AI, which would bias results if there are interactions between the AI and the reader specialty or anatomic location. Although the baseline condition (with AI or without AI) was randomized, and a washout period of at least 1 month was respected, there may be a carryover effect that would tend to disadvantage the better condition (ie, condition with AI); although we did not find one with statistical testing, the power of interaction tests is known to be poor. Third, our study used a consensus ground truth based on radiographic interpretation by expert musculoskeletal radiologists without CT. Overall, the context of readings was quite different from a real-life clinical setting, limiting the generalizability and clinical relevance of this work. Of note, use of AI does not improve detection of radiographically occult fracture because it is by definition "negative" on radiographs. Because the AI is designed to be used as an aid with human confirmation, any occult fracture that it would detect would likely be dismissed by the human reader. The inability of AI to detect radiographic occult fracture is not a limitation of the AI itself, but rather a limitation of radiography as a modality in general. Those fractures need to be depicted with cross-sectional imaging as clinically warranted.

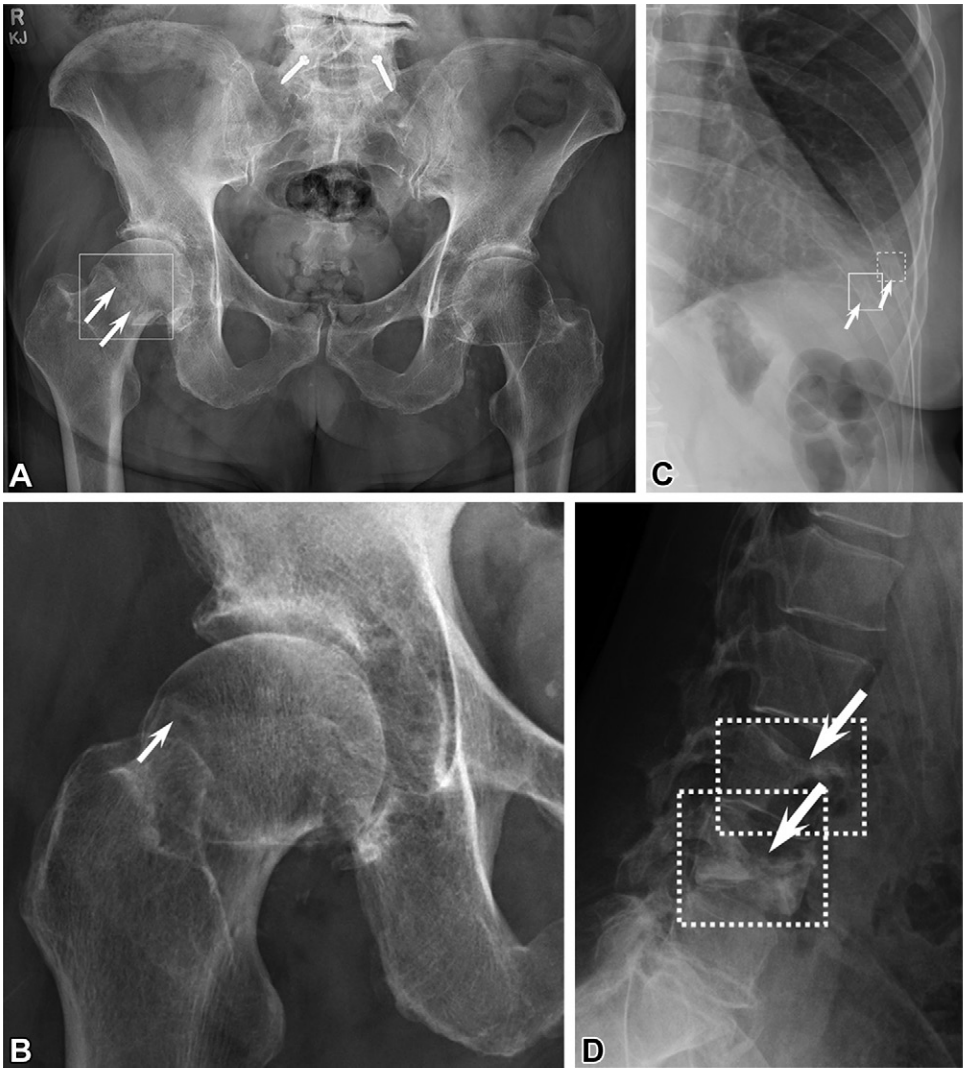

Figure 4: Stand-alone artificial intelligence (AI) performance examples: positive radiographs for fractures. **(A)** Radiograph shows a single true-positive fracture of the right femoral neck (arrows). This fracture was detected by AI using the FRACT threshold (box). One senior and one junior radiologist, two emergency department physicians, one physician assistant, three rheumatologists, and one family medicine physician missed the fracture. All readers pointed out the fracture with AI. **(B)** Additional dedicated view of right hip clearly shows this fracture (arrow). **(C)** Radiograph shows true-positive multiple left-sided rib fractures (arrows). One fracture was detected by AI using the FRACT threshold (solid box) and the other using the DOUBT-FRACT threshold (dashed box). Two senior and one junior radiologist, one senior orthopedic surgeon, one emergency department physician, and one physician assistant recognized the two rib fractures without AI. All readers pointed out the two rib fractures with AI. **(D)** Radiograph shows true-positive fractures of the L3 and L4 vertebral bodies (arrows). These fractures were detected by AI using the DOUBT-FRACT threshold (boxes). Thirteen readers pointed out the two fractures without AI. Nineteen readers pointed out the two fractures with AI. Two family medicine physicians, one rheumatologist, one radiology resident, and one physician assistant missed one vertebral fracture with and without AI. There were two predefined thresholds for fracture detection: high-sensitivity threshold named DOUBT-FRACT, equal to 50% after transformation, and high-specificity threshold named FRACT, equal to 90% after transformation.

In conclusion, radiographic artificial intelligence assistance improves the sensitivity, and may even improve the specificity, of fracture detection by radiologists and nonradiologists involving various anatomic locations. It also slightly reduces the time needed to interpret radiographs.

Author contributions: Guarantors of integrity of entire study, **A. Guermazi, A.M.M., A.D., X.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, **A. Guermazi, A.M.M., A.D., X.L., D.H.**; clinical studies, **A. Guermazi, A.J.K., A.M.M., A.D., A.T., Y.L., H.R., R.L.P., D.C., N.E.R., D.H.**; experimental studies, **A.D., Y.L., E.L., A.C.M., N.E.R.**; statistical analysis, **A. Gillibert, A.P.**; and manuscript editing, **A. Guermazi, C.T., A.J.K., A.M.M., A. Gillibert, X.L., Y.L., M.J., R.L.P., A.C.M., D.C., D.H.**

Disclosures of conflicts of interest: **A. Guermazi.** Consulting fees from Pfizer, TissueGene, MerckSerono, Novartis, Regeneron, AstraZeneca; stock options in BICL. **C.T.** Publishing royalties from Wolters-Kluwer; consulting fees from DePuy Synthes; Medical Director Orthopaedic Ambulatory Care – Boston University Medical Campus; stock options in 4web. **A.J.K.** No relevant relationships. **A.M.M.** No relevant relationships. **A.D.** No relevant relationships. **A. Gillibert.** No relevant relationships. **X.L.** No relevant relationships. **A.T.** No relevant relationships. **Y.L.** Advisory board for Regeneron; locum work to read x-rays for BICL; advisory board in osteoarthri-

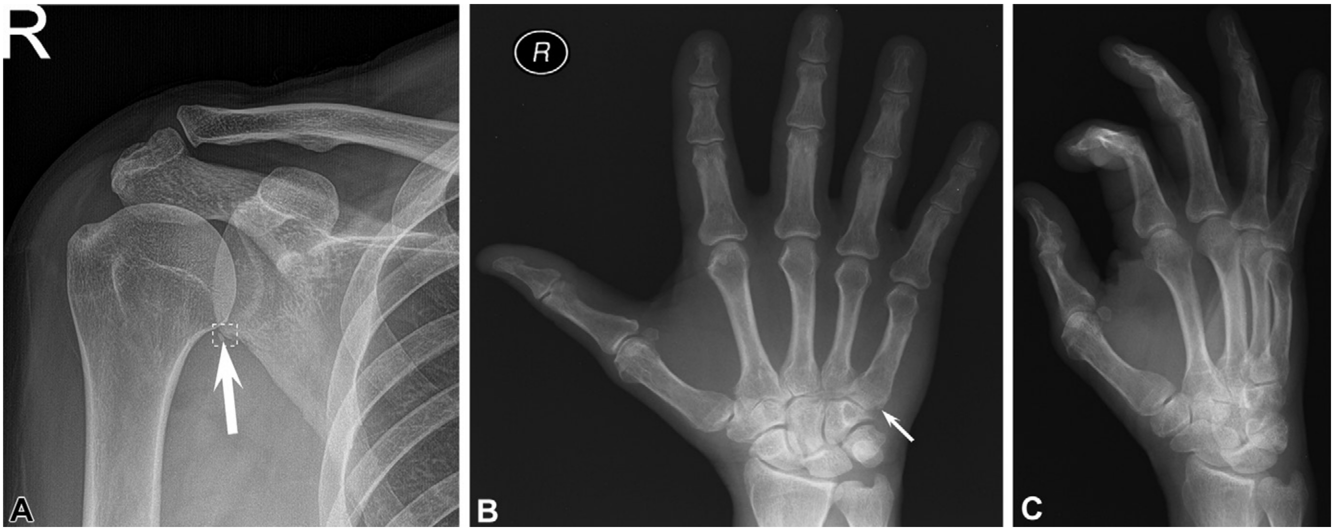

Figure 5: Stand-alone artificial intelligence (AI) performance examples: false-positive and false-negative radiographs. **(A)** Radiograph shows a small corticated ossific fragment adjacent to inferior glenoid margin (arrow), likely sequela of prior trauma (chronic fracture) or calcified detached inferior labrum rather than acute fracture. AI noted this as an acute fracture using the DOUBT-FRACT threshold. Fifteen readers read this as acute fracture without AI. Four readers thought the fracture was chronic without using AI, but reversed their reading with AI. Only two radiologists, one rheumatologist, and two family medicine physicians recognized the chronicity of the fracture with and without AI. **(B)** Radiograph shows a subtle nondisplaced fracture of the fifth metacarpal base (arrow), which was not detected by AI. All readers missed this fracture with and without AI. Only ground truth readers noted the fracture. This fracture was only appreciable on the anteroposterior view shown here and was not clearly visible on **(C)** the oblique view or the lateral view (not shown) of the right hand. There were two predefined thresholds for fracture detection: high-sensitivity threshold named DOUBT-FRACT, equal to 50% after transformation, and high-specificity threshold named FRACT, equal to 90% after transformation.

tis for Regeneron. **M.J.** Advisory board for Mass General Hospital; editorial board member for *Osteoarthritis Imaging*. **E.L.** No relevant relationships. **H.R.** No relevant relationships. **A.P.** No relevant relationships. **R.L.P.** Associate editor for *Journal of Cartilage & Joint Preservation*; editorial or governing board for *Arthroscopy*, *Arthroscopy, Sports Medicine and Rehabilitation*, and *Journal of Sport Rehabilitation*; board or committee member for American Orthopaedic Society for Sports Medicine and Society of Military Orthopaedic Surgeons; editorial reviewer for *American Journal of Sports Medicine*, *Orthopaedic Journal of Sports Medicine*, *Sports Health*, *The Bone and Joint Journal*, *Journal of Shoulder and Elbow Surgery*, *BMC Musculoskeletal Disorders*, and *The Orthopaedic Journal at Harvard Medical School*; Grant/Education and Hospitality Financial or Material Support from Arthrex. **A.C.M.** No relevant relationships. **D.C.** No relevant relationships. **N.E.R.** No relevant relationships. **D.H.** Editorial advisor for *BMC Musculoskeletal Disorders*; editorial board member for *Osteoarthritis Imaging*.

References

1. Gergenti L, Olympia RP. Etiology and disposition associated with radiology discrepancies on emergency department patients. Am J Emerg Med 2019;37(11):2015–2019.
2. Fernholm R, Pukk Härenstam K, Wachtler C, Nilsson GH, Holzmann MJ, Carlsson AC. Diagnostic errors reported in primary healthcare and emergency departments: A retrospective and descriptive cohort study of 4830 reported cases of preventable harm in Sweden. Eur J Gen Pract 2019;25(3):128–135.
3. Mattijssen-Horstink L, Langeraar JJ, Mauritz GJ, van der Stappen W, Baggelaar M, Tan ECTH. Radiologic discrepancies in diagnosis of fractures in a Dutch teaching emergency department: a retrospective analysis. Scand J Trauma Resusc Emerg Med 2020;28(1):38.
4. Fitschen-Oestern S, Lippross S, Lefering R, et al. Missed hand and forearm injuries in multiple trauma patients: An analysis from the TraumaRegister DGU®. Injury 2020;51(7):1608–1617.
5. Fitschen-Oestern S, Lippross S, Lefering R, et al. Missed foot fractures in multiple trauma patients. BMC Musculoskelet Disord 2019;20(1):121.
6. Tobler P, Cyriac J, Kovacs BK, et al. AI-based detection and classification of distal radius fractures using low-effort data labeling: evaluation of applicability and effect of training set size. Eur Radiol 2021;31(9):6816–6824.
7. Raisuddin AM, Vaattovaara E, Nevalainen M, et al. Critical evaluation of deep neural networks for wrist fracture detection. Sci Rep 2021;11(1):6006.
8. Kim DH, MacKinnon T. Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks. Clin Radiol 2018;73(5):439–445.
9. Ren M, Yi PH. Deep learning detection of subtle fractures using staged algorithms to mimic radiologist search pattern. Skeletal Radiol 2021. https://doi.org/10.1007/s00256-021-03739-2. Published online February 12, 2021.
10. Cheng CT, Ho TY, Lee TY, et al. Application of a deep learning algorithm for detection and visualization of hip fractures on plain pelvic radiographs. Eur Radiol 2019;29(10):5469–5477.
11. Cheng CT, Wang Y, Chen HW, et al. A scalable physician-level deep learning algorithm detects universal trauma on pelvic radiographs. Nat Commun 2021;12(1):1066.
12. Chen HY, Hsu BW, Yin YK, et al. Application of deep learning algorithm to detect and visualize vertebral fractures on plain frontal radiographs. PLoS One 2021;16(1):e0245992.
13. Lindsey R, Daluiski A, Chopra S, et al. Deep neural network improves fracture detection by clinicians. Proc Natl Acad Sci U S A 2018;115(45):11591–11596.
14. Jones RM, Sharma A, Hotchkiss R, et al. Assessment of a deep-learning system for fracture detection in musculoskeletal radiographs. NPJ Digit Med 2020;3(1):144.
15. Wu Y. Kirillov A, Massa F, Lo WY, Girshick R. Detectron2. https://github.com/facebookresearch/detectron2. Published 2019. Accessed March 1, 2020.
16. Duron L, Ducarouge A, Gillibert A, et al. Assessment of an AI aid in detection of adult appendicular skeletal fractures by emergency physicians and radiologists: a multicenter cross-sectional diagnostic study. Radiology 2021;300(1):120–129.
17. Lee CS, Nagy PG, Weaver SJ, Newman-Toker DE. Cognitive and system factors contributing to diagnostic errors in radiology. AJR Am J Roentgenol 2013;201(3):611–617.
18. Hartigan S, Brooks M, Hartley S, Miller RE, Santen SA, Hemphill RR. Review of the basics of cognitive error in emergency medicine: Still no easy answers. West J Emerg Med 2020;21(6):125–131.
19. Krupinski EA, Berbaum KS, Caldwell RT, Schartz KM, Kim J. Long radiology workdays reduce detection and accommodation accuracy. J Am Coll Radiol 2010;7(9):698–704.
20. Reiner BI, Krupinski E. The insidious problem of fatigue in medical imaging practice. J Digit Imaging 2012;25(1):3–6.
21. Gaba DM, Howard SK. Patient safety: fatigue among clinicians and the safety of patients. N Engl J Med 2002;347(16):1249–1255.
22. Dembrower K, Wahlin E, Liu Y, et al. Effect of artificial intelligence-based triaging of breast cancer screening mammograms on cancer detection and radiologist workload: a retrospective simulation study. Lancet Digit Health 2020;2(9):e468–e474.
23. Castillo C, Steffens T, Sim L, Caffery L. The effect of clinical information on radiology reporting: A systematic review. J Med Radiat Sci 2021;68(1):60–74.
24. Egglin TKP, Feinstein AR. Context bias. A problem in diagnostic radiology. JAMA 1996;276(21):1752–1755.