Radiology

ORIGINAL RESEARCH • MUSCULOSKELETAL IMAGING

Assessment of an AI Aid in Detection of Adult Appendicular Skeletal Fractures by Emergency Physicians and Radiologists: A Multicenter Cross-sectional Diagnostic Study

Loïc Duron, MD, MSc • Alexis Ducarouge, MSc • André Gillibert, MD, MSc • Julia Lainé, MD, MSc • Christian Allouche • Nicolas Cherel, MSc • Zekun Zhang, MSc • Nịcolas Nitche, MSc • Elise Lacave, MSc • Aloïs Pourchot, MSc • Adrien Felter, MD • Louis Lassalle, MD, MSc • Nor-Eddine Regnard, MD, MSc • Antoine Feydy, MD, PhD

From the Department of Radiology, Hôpital Fondation A. de Rothschild, 25 rue Manin, 75019 Paris, France (L.D.); Faculty of Medicine, Université de Paris, Paris, France (L.D., A. Feydy); Gleamer, Paris, France (A.D., C.A., N.C., Z.Z., N.N., E.L., A.P., N.E.R.); Department of Biostatistics, CHU Rouen, Rouen, France (A.G.); Department of Radiology, Hôpital Hôtel-Dieu, Assistance Publique-Hôpitaux de Paris, Paris, France (J.L.); Department of Radiology, Hôpital Ambroise-Paré, Assistance Publique-Hôpitaux de Paris, Boulogne-Billancourt, France (A. Felter); Department of Radiology, Hôpital Raymond-Poincaré, Assistance Publique-Hôpitaux de Paris, Garches, France (A. Felter); and Department of Radiology B, Hôpital Cochin, Assistance Publique-Hôpitaux de Paris, Paris, France (L.L., N.E.R., A. Feydy). Received September 30, 2020; revision requested December 23; revision received January 26, 2021; accepted March 4. **Address correspondence to** L.D. (e-mail: *lduron@for.paris*).

This study was funded by Gleamer.

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

Radiology 2021; 000:1–10 • https://doi.org/10.1148/radiol.2021203886 • Content codes: MK AI

Background: The interpretation of radiographs suffers from an ever-increasing workload in emergency and radiology departments, while missed fractures represent up to 80% of diagnostic errors in the emergency department.

Purpose: To assess the performance of an artificial intelligence (AI) system designed to aid radiologists and emergency physicians in the detection and localization of appendicular skeletal fractures.

Materials and Methods: The AI system was previously trained on 60 170 radiographs obtained in patients with trauma. The radiographs were randomly split into 70% training, 10% validation, and 20% test sets. Between 2016 and 2018, 600 adult patients in whom multiview radiographs had been obtained after a recent trauma, with or without one or more fractures of shoulder, arm, hand, pelvis, leg, and foot, were retrospectively included from 17 French medical centers. Radiographs with quality precluding human interpretation or containing only obvious fractures were excluded. Six radiologists and six emergency physicians were asked to detect and localize fractures with (n = 300) and fractures without (n = 300) the aid of software highlighting boxes around AI-detected fractures. Aided and unaided sensitivity, specificity, and reading times were compared by means of paired Student t tests after averaging of performances of each reader.

Results: A total of 600 patients (mean age ± standard deviation, 57 years ± 22; 358 women) were included. The AI aid improved the sensitivity of physicians by 8.7% (95% CI: 3.1, 14.2; P = .003 for superiority) and the specificity by 4.1% (95% CI: 0.5, 7.7; P < .001 for noninferiority) and reduced the average number of false-positive fractures per patient by 41.9% (95% CI: 12.8, 61.3; P = .02) in patients without fractures and the mean reading time by 15.0% (95% CI: −30.4, 3.8; P = .12). Finally, stand-alone performance of a newer release of the AI system was greater than that of all unaided readers, including skeletal expert radiologists, with an area under the receiver operating characteristic curve of 0.94 (95% CI: 0.92, 0.96).

Conclusion: The artificial intelligence aid provided a gain of sensitivity (8.7% increase) and specificity (4.1% increase) without loss of reading speed.

Online supplemental material is available for this article.

Traumatic skeletal injuries are a leading source of consultation in emergency departments, with an annual incidence reported to be as high as 1.3% in the United States (1) and 0.32% in China (2). Radiography is the first-line imaging modality for the diagnosis of these lesions and the most used imaging modality worldwide (3–5). The reading of trauma radiographs is a demanding task that requires radiologic expertise, and there is a lack of radiologists (6). Consequently, emergency physicians are required to make patient treatment decisions before the availability of a radiologist's report, with a risk of interpretation error (7–9). Missed fractures, a preventable cause of morbidity (10), represent up to 80% of emergency department diagnostic errors (11). In American medical-legal claims, extremity fractures are the second most frequently missed diagnosis leading to a claim, after breast cancer (12). Assisting physicians in detecting and localizing fractures on plain radiographs could therefore reduce error rates.

Computer-aided detection software has been developed for more than 20 years to provide decision support to radiologists, especially for screening breast cancer on mammograms (13) and lung nodules on CT scans (14). However, computer-aided detection systems have a high false-positive rate, which has limited their acceptance (13). Similar technologies have been unsuccessfully investigated for fracture detection, potentially because of

Abbreviations

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

Summary

The artificial intelligence aid improved the sensitivity and specificity of radiologists and emergency physicians in the localization of appendicular fractures on radiographs, with no additional reading time.

Key Results

- The artificial intelligence (AI) aid, which highlighted potential fractures on full-resolution radiographs, improved the sensitivity (8.7% increase, *P* = .006) and specificity (4.1% increase, *P* = .03) of emergency doctors and radiologists in the diagnosis of appendicular fractures.
- The stand-alone area under the receiver operating characteristic curve, requiring that the AI system detect the precise locations of all fractures on an examination, was .94 with a newer release of the AI system.

the wider variety of morphologic aspects and image patterns involved (15,16). Only the recent advent of deep learning, which has helped reach superhuman performances in a wide variety of image-related tasks (17,18), has allowed the emergence of systems capable of tackling this challenge. Several recent studies have assessed the performance of algorithms in the detection of bone fractures, with encouraging results (19–29). However, all of these approaches have limitations that jeopardize their applicability in a practical context. For instance, to our knowledge, no algorithms have been proposed to analyze all body parts simultaneously or to detect multiple fractures in a single radiograph, which are frequently encountered in every clinical practice.

In this study, we appraised the performance of a new artificial intelligence (AI) system trained to detect and localize all limb and pelvis fractures. By using a data set of posttraumatic radiographs from 17 centers, we compared the detection performance and reading time, with and without AI aid, in a clinically relevant setup of emergency physicians and radiologists of various levels of experience. We also assessed the stand-alone AI system performance.

Materials and Methods

This 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.G., a biostatistician with 4 years of experience, and L.G., a radiologist with 7 years of experience). A.D., C.A., N.E.R., N.C., Z.Z., N.N, E.L., and A.P. are or were employees of Gleamer. A. Feydy is a paid consultant for Gleamer. A. Felter and L.L. were paid by Gleamer to participate in the study as readers or experts. L.D., A.G., and J.L. have no conflicts of interest to declare.

This retrospective study was approved by the institutional review board, which waived specific patient informed consent. We followed the Standards for Reporting of Diagnostic Accuracy Studies guidelines and report key considerations of radiology AI studies (30) in Table E1 (online).

Model Building and Validation

We gathered a development data set of 60 170 radiographs in patients with trauma from 22 French public hospitals and private radiology departments from January 2011 to May 2019; this data set was randomly split into 70% training, 10% validation,

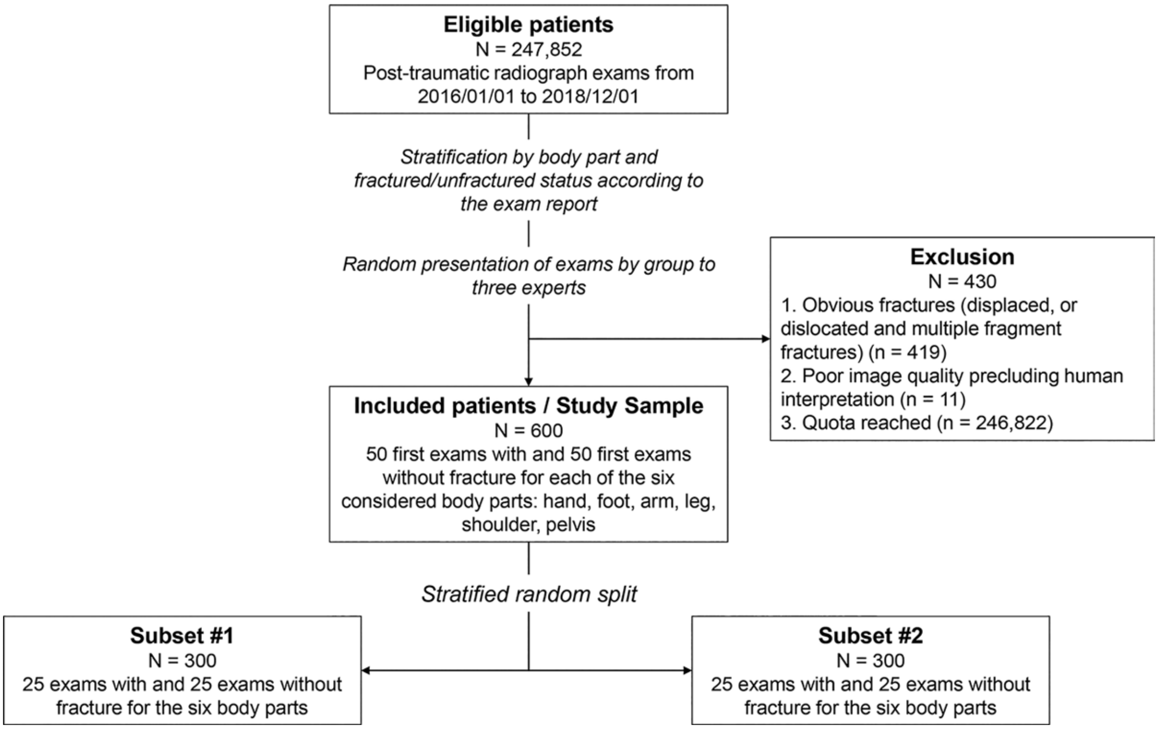
Eligible patients
N = 247,852
Post-traumatic radiograph exams from 2016/01/01 to 2018/12/01
Stratification by body part and fractured/unfractured status according to the exam report
Random presentation of exams by group to three experts
Exclusion
N = 430
1. Obvious fractures (displaced, or dislocated and multiple fragment fractures) (n = 419)
2. Poor image quality precluding human interpretation (n = 11)
3. Quota reached (n = 246,822)
Included patients / Study Sample
N = 600
50 first exams with and 50 first exams without fracture for each of the six considered body parts: hand, foot, arm, leg, shoulder, pelvis
Stratified random split
Subset #1
N = 300
25 exams with and 25 exams without fracture for the six body parts
Subset #2
N = 300
25 exams with and 25 exams without fracture for the six body parts

Figure 1: Flowchart of study.

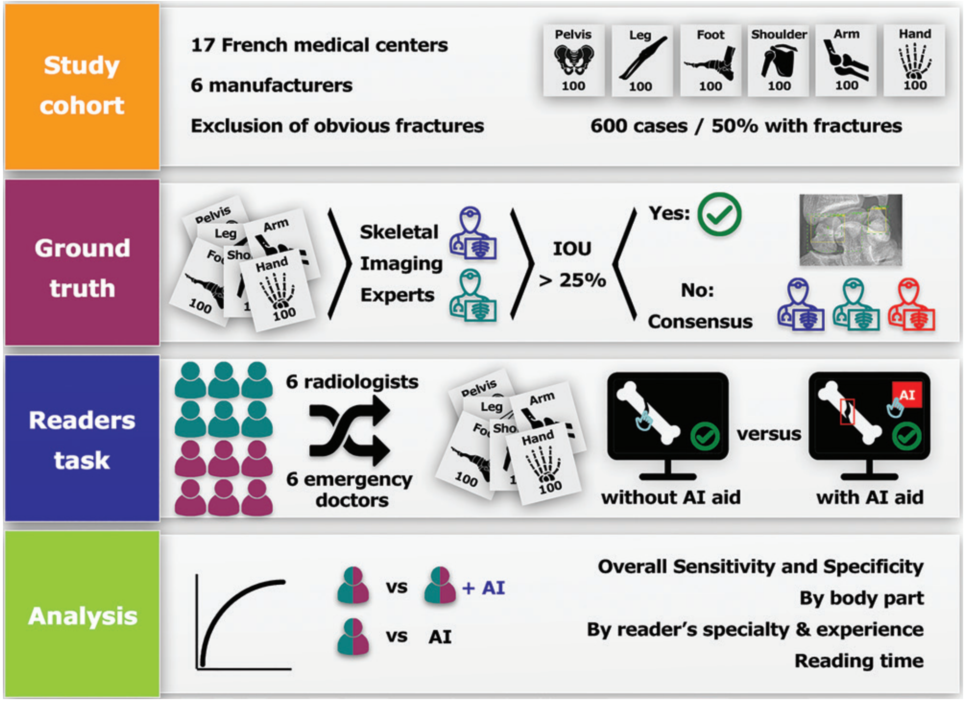

Figure 2: Diagram depicts study design. Each step is illustrated in successive boxes, including study sample constitution, establishment of ground truth by three experts in skeletal imaging, reading tasks with and without artificial intelligence (AI) aid, and main lines of data analysis. IOU = intersection over union.

and 20% internal test sets. A deep convolutional neural network based on the "Detectron 2" (31) framework was engineered, trained, optimized, and validated to detect and localize fractures on native resolution digital radiographs. The AI system was integrated into radiology software as a diagnostic aid, highlighting each region of interest with a box and providing a confidence score regarding the existence of a fracture in the region of interest (Appendix E1 [online]).

Inclusion and Exclusion Criteria

The study sample, an external test data set, was retrospectively built from examinations (sets of radiographs taken at the same time in a patient) completed between January 2016 and December 2017 in 17 other French medical centers with use of digital radiography systems from three manufacturers found in the development data set (Philips Medical Systems, PrimaX, ATS) and three other manufacturers (IDETEC Digital Medical Imaging, FujiFilm, Siemens). No study sample radiograph was used in the development data set.

Inclusion criteria were age 18 years or older and at least one digital plain radiograph of an appendicular body part (shoulder, arm, hand, pelvis, leg, foot) obtained after a recent trauma, with or without fracture.

Exclusion criteria were *(a)* poor radiographic quality precluding human interpretation, *(b)* examinations showing only obvious fractures (displaced, dislocated, or multiple fragments) according to ground-truth radiologists, and *(c)* quota of 50 patients reached for that body part and fracture status (fractured vs unfractured).

Stratified randomized sampling was used to include 50 patients with and 50 patients without fracture for each location (Appendix E1 [online]), leading to 2441 radiographs in 600 examinations of 600 patients (Figs 1, 2). The sample was split once into two 300-patient subsets with stratified randomization so that the resulting subsets were similar in terms of median age, female-to-male ratio, body part, and fracture prevalence.

Ground Truth

Two skeletal imaging radiologists (L.L. and N.E.R., with 9 and 10 years of experience, respectively) independently drew bounding boxes around each bone fracture they detected in each view of the 600 examinations by using dedicated software, without AI aid. The ground truth of a fracture was defined as the union of the experts' bounding boxes when their bounding boxes had an intersection over union above 25%. Disagreements were resolved (149 examinations) by majority consensus with a third skeletal imaging radiologist (A. Feydy, with 28 years of experience).

Readers and Readings

Twelve independent readers (six radiologists and six emergency physicians) of various levels of experience (including residents and experts) working in different French radiology centers were enrolled from June to August 2019 and trained to the detection task with AI aid on a small independent sample of cases. Readers were blinded to clinical data, and readings were performed on dedicated medical workstations displaying de-identified lossless radiographs, without any time constraints. Reading time of each examination was automatically recorded. Readers localized fractures by pinpointing them on at least one radiograph of a displayed patient case. Readers and the two 300-patient subsets were randomized so that half of the readers interpreted subset 1 with AI aid and subset 2 without AI aid, while the other half interpreted subset 1 without AI aid and subset 2 with AI aid, and an even distribution of specialty and experience was respected (Table E2 [online]). Readers were blinded to one another and to expert's judgments.

Metrics

Patient-wise classification metrics.—Patient-wise sensitivity was defined as the proportion of examinations in which all actual fractures were discovered and indicated (each one, on at least one radiograph) among examinations having at least one fracture, even if some incorrect spots (false-positive findings) had been added by the reader. Patient-wise specificity was defined as the proportion of examinations in which no fracture was spotted by the reader among examinations having no fracture.

Fracture-wise object detection metrics.—Patient-wise definitions ignore the partial detection of multiple fractures and false-positive fractures when all real fractures have been indicated. To overcome these limitations, the fracture-wise sensitivity was defined as the proportion of fractures correctly detected among all fractures, on at least one radiograph with observations being fractures rather than examinations. The average number of false-positive fractures per patient was defined as the average number of spots per examination placed at regions where there was no fracture.

Performance of Readers

The two co-primary outcomes were patient-wise sensitivity and patient-wise specificity. The primary analysis had to demonstrate both the superiority of the average patient-wise sensitivity with aid to that without aid and the noninferiority of the average patient-wise specificity with aid to that without aid with a noninferiority margin of −3%.

Subgroup Analyses

Readers results were also analyzed according to reader's specialty, fracture location, fracture severity (as defined in Appendix E1 [online]), and number of fractures per patient.

Stand-alone AI Performance

The stand-alone algorithm performance was assessed by using receiver operating characteristic and free-response receiver operating characteristic curves based on the same definitions of patient-wise and fracture-wise diagnosis performances, respectively. To be consistent with the evaluation of human readers, who used a dot to mark the area of fracture, we considered the center of the region of interest as the location of the AI-indicated fractures. During the study a new version of the AI system was developed owing to better annotations of training data, but because the AI system used in the diagnostic aid software was unchanged, performances of AI-aided readers were based on the older AI system. Stand-alone performances of both the older and the newer versions of the AI system were assessed by using the study sample. A post hoc sensitivity analysis considering the diagnosis of each patient as binary (fracture[s] vs no fracture) was also performed.

Statistical Analysis

Statistical analyses were performed by A.G. using statistical software (R, version 4.0.2; R Foundation for Statistical Computing).

Primary analysis.—For each reader, the mean patient-wise sensitivity and specificity with and without aid were computed and then compared between aided and unaided readings by means of paired *t* tests with 12 pairs of observations (one pair for each reader). The success of primary analysis required both the superiority of sensitivity and noninferiority of specificity with a noninferiority margin of −3%, both at 2.5% one-sided significance threshold. All other tests were two sided at 5% significance threshold.

Secondary analyses.—Patient-wise sensitivity and specificity, fracture-wise sensitivity, average number of false-positive fractures per patient, and reading time were averaged for each reader, with or without aid, then compared by means of appropriate paired or two-sample *t* tests or general linear models. Receiver operating characteristic and free-response receiver operating characteristic curves were drawn by using the custom metrics (patient-wise sensitivity and specificity, fracture-wise sensitivity, and average number of false-positive fractures per patient). For comparison with the literature, the usual binary definition of fracture (yes or no) was used in a sensitivity analysis. Statistical analyses, including subgroup analyses, are further described in Appendix E1 (online).

Results

Patient Characteristics

During the screening, 419 examinations that contained only obvious fractures and 11 with poor radiographic quality precluding human interpretation were excluded (Fig 1). The mean age of the 600 included patients (±standard deviation) was 57 years ± 22 (range, 18–100 years); 358 of the 600 patients (60%) were women. Women were older than men (mean age, 62 years vs 49 years, respectively; $P < .001$), and 229 of the 358 women (64%) were aged 55 years or older. Patients with at least one fracture were older than those without fractures (mean age, 62 years vs 51 years, respectively; $P < .001$). A summary of demographic characteristics according to location and fracture status is given in Table 1. Fracture sites are listed in Table E3 (online).

Performance of Readers

The use of AI to read radiographic examinations improved the patient-wise sensitivity from 70.8% to 79.4% (8.7% increase; 95% CI: 3.1, 14.2; $P = .003$ for superiority) and patient-wise specificity from 89.5% to 93.6% (4.1% increase; 95% CI: 0.5,

Table 1: Summary of Demographic Characteristics according to Examination Location and Presence of Fracture

	Patients with Fracture		Patients without Fracture	
Location	Age (y)*	No. of Women	Age (y)*	No. of Women
Total	62 ± 22	192/300 (64)	51 ± 21	166/300 (55)
Shoulder	68 ± 19	26/42 (62)	55 ± 19	18/37 (49)
Arm	51 ± 20	26/37 (70)	48 ± 15	19/45 (42)
Hand	51 ± 25	23/44 (52)	40 ± 17	28/50 (56)
Pelvis	82 ± 10	34/45 (76)	66 ± 24	27/46 (59)
Leg	60 ± 20	30/46 (65)	52 ± 18	22/44 (50)
Foot	59 ± 17	27/44 (61)	47 ± 20	27/40 (68)
Multiple locations	59 ± 24	26/42 (62)	53 ± 21	25/38 (66)

Note.—Unless otherwise specified, data are numbers of patients with percentages in parentheses.

* Numbers are means ± standard deviations.

Table 2: Diagnostic Performances of Unaided and Aided Readings

Parameter	Unaided (n = 12)*	Aided (n = 12)*	Absolute Difference (n = 12)†	Relative Difference (%) (n = 12)†
SE$_{PW}$ (%)	70.8 ± 12.5	79.4 ± 7.4	+8.7 (3.1, 14.2) [.003]‡	12.2 (4.0, 21.2)
SPE$_{PW}$ (%)	89.5 ± 6.5	93.6 ± 4.6	+4.1 (0.5, 7.7) [<.001]‡	4.6 (0.7, 8.6)
SE$_{FW}$ (%)	73.7 ± 11.1	81.2 ± 6.5	+7.5 (2.8, 12.2) [.005]	10.2 (3.5, 17.2)
PP-FP$_{FW}$ for patients without fracture	0.113 ± 0.069	0.066 ± 0.048	−0.047 (−0.086, −0.009) [.02]	−41.9 (−61.3, −12.8)
PP-FP$_{FW}$ for patients with fracture	0.082 ± 0.055	0.045 ± 0.028	−0.037 (−0.073, 0.000) [.05]	−44.9 (−67.0, −7.9)
Mean reading time (sec)	67.0 ± 26.2	57.0 ± 24.8	−10.0 (−23.1, 3.0) [.12]	−15.0 (−30.4, 3.8)

Note.—PP-FP$_{FW}$ = average number of false-positive fractures per patient, SE$_{FW}$ = fracture-wise sensitivity, SE$_{PW}$ = patient-wise sensitivity, SPE$_{PW}$ = patient-wise specificity,

* Data are means ± standard deviations.

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

‡ P values are one-sided for primary analysis with superiority margin of +0% for SE$_{PW}$ and noninferiority margin of −3% for SPE$_{PW}$.

7.7; P < . 001 for noninferiority). The mean average number of false-positive fractures per patient in patients with no fractures was reduced from 0.113 to 0.066 (41.9% decrease; 95% CI: 12.8, 61.3; P = .02). The mean reading time was reduced from 67.0 to 57.0 seconds (10.0-second decrease; 95% CI: −3.0, 23.1; P = .12) with AI aid. AI aid improved the fracture-wise sensitivity (absolute proportion difference) by 7.5% (95% CI: 2.8, 12.2; P = .005). Results of sensitivity analyses, shown in Appendix E2 (online), were consistent with the primary analysis. Table 2 and Figure 3 show changes in metrics with AI aid. Table E4 (online) shows raw reader performances.

Subgroup Analyses

Tables 3 and E5 (online) and Appendix E1 (online) show results of subgroup analyses. A gain in patient-wise sensitivity was observed with AI aid for all body parts; however, a statistically significant difference was seen only for the hand and foot.

The sensitivity of emergency physicians improved from 61.3% to 74.3% (13.0% increase, P = .03) with AI, and the sensitivity of radiologists improved from 80.2% to 84.6% (4.3% increase, P = .03). The difference in sensitivity gain between emergency physicians (13.0% increase) and radiologists (4.3% increase) was estimated at 8.7% (95% CI: −1.4, 18.7; P = .08). After adjustment for the unaided sensitivity (61.3% on average for emergency physicians vs 80.2% for radiologists), this difference in sensitivity gain was reduced to −5.6% (95% CI: −16.7, 5.4; P = .28) for emergency physicians versus radiologists.

The patient-wise sensitivity of aided emergency physicians (74.3%) was 5.9% (95% CI: −2.6, 14.4; P = .15) lower than that of unaided physicians (80.2%), whereas the patient-wise specificity of aided emergency physicians (96.6%) was higher than that of unaided radiologists (88.4%) (absolute difference, 8.1%; 95% CI: 1.1, 15.2; P = .03). Overall, the absolute difference of Youden index (patient-wise sensitivity + patient-wise specificity − 1) between aided emergency physicians and unaided radiologists was estimated at +2.2% (95% CI: −5.1, 9.6; P = .51).

The difference in average fracture-wise sensitivity gains between the 99 patients with severe fractures (absolute gain, 6.0%; 95% CI: −0.3, 12.4; P = .06) and the 201 patients with nonsevere fractures (absolute gain, 9.9%; 95% CI: 4.3, 15.5; P = .003) was estimated at −3.8% (95% CI: −8.0, 0.3; P = .07). Gains in fracture-wise sensitivity were not significantly different (absolute difference, −3.1%; 95% CI: −8.2, 2.1; P = .22) between the 59 patients with multiple fractures (absolute difference, 5.5%; 95% CI: 1.3, 9.7; P = .01) and the 241 patients

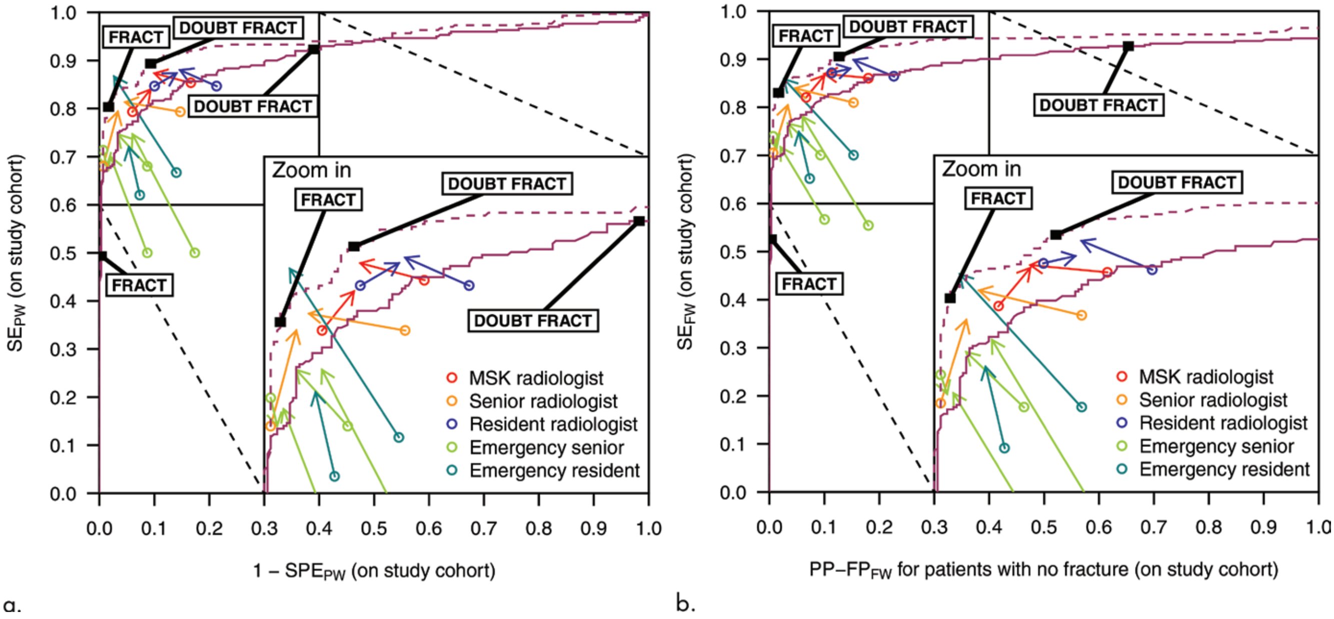

Figure 3: Receiver operating characteristic and free-response receiver operating characteristic curves show artificial intelligence (AI) performance and unaided and aided reader performance. **(a)** Receiver operating characteristic curves and **(b)** free-response receiver operating characteristic curves show the stand-alone performance of the older (solid line) and newer (dashed line) versions of the algorithm and the performances of radiologists with (ends of arrows) and without (circles) the aid of the AI system. A high-sensitivity threshold and a high-specificity threshold (detailed in Appendix E1 [online]) are used by the AI aid to respectively highlight possible fractures (DOUBT FRACT) and definite fractures (FRACT) as shown in Figures 4 and 5. Readers are grouped according to specialty. AI aid of readers was based on the older version of the algorithm, as the newer version was developed during the study. All curves are computed for the study sample. AUC = area under the receiver operating characteristic curve, MSK = musculoskeletal specialist, PP-FP_{FW} = average number of false-positive fractures per patient, SE_{FW} = fracture-wise sensitivity, SE_{PW} = patient-wise sensitivity, SPE_{PW} = patient-wise specificity

Table 3: Predictive Performances of the 12 Readers with and without AI Aid according to Body Part

Body Part and Parameter	Unaided*	Aided*	Absolute Difference†
Shoulder			
SE_{PW} (%)	75.3 ± 17.2	80.9 ± 10.9	5.5 (−2.3, 13.3) [.15]
SPE_{PW} (%)	97.6 ± 3.0	92.8 ± 9.7	−4.9 (−10.0, 0.3) [.06]
Y_{PW} (%)	73.0 ± 16.6	73.7 ± 10.0	0.7 (−10.0, 11.3) [.89]
SE_{FW} (%)	77.0 ± 16.2	82.1 ± 10.6	5.0 (−2.3, 12.4) [.16]
PP-FP_{FW} for patients without fracture	0.029 ± 0.040	0.072 ± 0.097	0.043 (−0.005, 0.092) [.08]
PP-FP_{FW} for patients with fracture	0.046 ± 0.061	0.054 ± 0.077	0.008 (−0.059, 0.074) [.81]
Mean reading time (sec)	52.7 ± 20.1	54.9 ± 28.5	2.2 (−15.1, 19.5) [.78]
Arm			
SE_{PW} (%)	83.2 ± 16.9	86.7 ± 9.7	3.5 (−9.0, 16.1]) [.55]
SPE_{PW} (%)	93.8 ± 5.8	97.0 ± 3.5	3.2 (0.2, 6.2) [.04]
Y_{PW} (%)	76.9 ± 18.1	83.7 ± 8.1	6.8 (−7.0, 20.5) [.30]
SE_{FW} (%)	83.5 ± 16.5	86.9 ± 9.8	3.4 (−9.2, 15.9) [.56]
PP-FP_{FW} for patients without fracture	0.062 ± 0.058	0.030 ± 0.035	−0.032 (−0.062, −0.002) [.04]
PP-FP_{FW} for patients with fracture	0.055 ± 0.063	0.042 ± 0.075	−0.013 (−0.081, 0.055) [.67]
Mean reading time (sec)	49.3 ± 19.5	43.3 ± 20.7	−6.0 (−14.9, 2.9) [.17]
Hand			
SE_{PW} (%)	59.6 ± 20.5	80.2 ± 11.4	20.5 (6.3, 34.8) [.009]
SPE_{PW} (%)	84.7 ± 11.0	91.0 ± 6.4	6.3 (−1.0, 13.6) [.08]
Y_{PW} (%)	44.3 ± 26.1	71.2 ± 8.7	26.9 (7.8, 46.0) [.01]
SE_{FW} (%)	66.4 ± 17.0	80.9 ± 9.4	14.6 (2.6, 26.5) [.02]
PP-FP_{FW} for patients without fracture	0.160 ± 0.119	0.090 ± 0.064	−0.070 (−0.150, 0.010) [.08]

Table 3 *(continues)*

Table 3 *(continued)*: Predictive Performances of the 12 Readers with and without AI Aid according to Body Part

Body Part and Parameter	Unaided*	Aided*	Absolute Difference[†]
PP-FP_{FW} for patients with fracture	0.118 ± 0.105	0.022 ± 0.03	−0.095 (−0.164, −0.027) [.01]
Mean reading time (sec)	83.2 ± 37.6	61.6 ± 29.5	−21.7 (−38.4, −4.9) [.02]
Pelvis			
SE_{PW} (%)	64.8 ± 12.4	69.8 ± 7.3	5.0 (−2.6, 12.6) [.18]
SPE_{PW} (%)	86.6 ± 9.1	91.8 ± 8.1	5.1 (−3.3, 13.6) [.21]
Y_{PW} (%)	51.5 ± 12.7	61.6 ± 9.5	10.1 (−2.5, 22.7) [.11]
SE_{FW} (%)	70.0 ± 10.8	73.8 ± 6.5	3.7 (−1.6, 9.1) [.15]
PP-FP_{FW} for patients without fracture	0.158 ± 0.114	0.082 ± 0.081	−0.076 (−0.174, 0.022) [.12]
PP-FP_{FW} for patients with fracture	0.091 ± 0.092	0.077 ± 0.084	−0.013 (−0.082, 0.055) [.67]
Mean reading time (sec)	59.1 ± 21.1	54.2 ± 25.2	−4.9 (−21.3, 11.5) [.53]
Leg			
SE_{PW} (%)	76.4 ± 11.8	80.1 ± 13.2	3.6 (−5.3, 12.5) [.39]
SPE_{PW} (%)	88.6 ± 10.5	96.2 ± 4.3	7.6 (0.9, 14.2) [.03]
Y_{PW} (%)	65.1 ± 13.9	76.3 ± 12.9	11.2 (1.5, 20.9) [.03]
SE_{FW} (%)	76.3 ± 11.8	80.4 ± 13.5	4.1 (−5.3, 13.5) [.36]
PP-FP_{FW} for patients without fracture	0.117 ± 0.115	0.042 ± 0.049	−0.076 (−0.150, −0.002) [.046]
PP-FP_{FW} for patients with fracture	0.051 ± 0.052	0.011 ± 0.027	−0.040 (−0.080, 0.000) [.05]
Mean reading time (sec)	57.0 ± 24.3	48.5 ± 19.4	−8.5 (−21.5, 4.6) [.18]
Foot			
SE_{PW} (%)	71.8 ± 13.6	86.9 ± 8.3	15.1 (7.5, 22.8) [.001]
SPE_{PW} (%)	88.0 ± 9.9	92.9 ± 5.8	4.9 (0.2, 9.5) [.04]
Y_{PW} (%)	59.8 ± 13.3	79.8 ± 7.8	20.0 (11.3, 28.7) [<.001]
SE_{FW} (%)	78.1 ± 10.8	90.0 ± 6.0	12.0 (6.2, 17.7) [<.001]
PP-FP_{FW} for patients without fracture	0.128 ± 0.112	0.071 ± 0.058	−0.057 (−0.110, −0.005) [.03]
PP-FP_{FW} for patients with fracture	0.102 ± 0.108	0.049 ± 0.036	−0.053 (−0.108, 0.003) [.06]
Mean reading time (sec)	76.3 ± 30.1	60.6 ± 28.3	−15.8 (−28.8, −2.8) [.02]
Multiple locations			
SE_{PW} (%)	65.6 ± 13.4	72.4 ± 10.7	6.8 (−2.5, 16.1) [.14]
SPE_{PW} (%)	88.2 ± 14.7	94.5 ± 6.7	6.3 (−0.3, 12.9) [.06]
Y_{PW} (%)	53.9 ± 16.8	67.0 ± 10.3	13.1 (0.9, 25.3) [.04]
SE_{FW} (%)	69.2 ± 10.3	77.3 ± 9.7	8.0 (−0.2, 16.2) [.05]
PP-FP_{FW} for patients without fracture	0.123 ± 0.147	0.062 ± 0.080	−0.061 (−0.120, −0.002) [.04]
PP-FP_{FW} for patients with fracture	0.112 ± 0.118	0.066 ± 0.054	−0.047 (−0.140, 0.047) [.30]
Mean reading time (sec)	90.3 ± 36.8	77.0 ± 31.6	−13.3 (−29.7, 3.1) [.10]

Note.—Performance was compared with ground truth. AI = artificial intelligence, PP-FP_{FW} = average number of false-positive fractures per patient, SE_{FW} = fracture-wise sensitivity, SE_{PW} = patient-wise sensitivity, SPE_{PW} = patient-wise specificity, Y_{PW} = Youden index by patient.

* Unless otherwise specified, data are means ± standard deviations.

[†] Absolute mean difference of performance attributed to AI aid, estimated with the Student paired *t* test in the 12 readers (12 observations). No multiple testing procedures were applied. Numbers in parentheses are the 95% CIs. Numbers in brackets are *P* values.

with a single fracture (absolute difference, 8.6%; 95% CI: 2.8, 14.4; *P* = .008).

Examples of fractures are illustrated in Figures 4 and 5.

Stand-alone AI Performance

The stand-alone area under the receiver operating characteristic curve (AUC) of the AI system study version was .91 (95% CI: .89, .94; *P* < .001). A newer version of the AI system, developed during the study but not included in the software, outperformed every unaided reader, showing an AUC of .94 (95% CI: .92, .96; *P* < .001). In a post hoc sensitivity analysis with a binary diagnosis for each patient (fracture vs no fracture), the AUCs of the older and newer AI system versions were, respectively, .95 (95% CI: .93, .97; *P* < .001) and .97 (95% CI: .96, .98; *P* < .001).

Discussion

Fractures represent up to 80% of emergency department diagnostic errors. Assisting physicians in detecting and localizing fractures

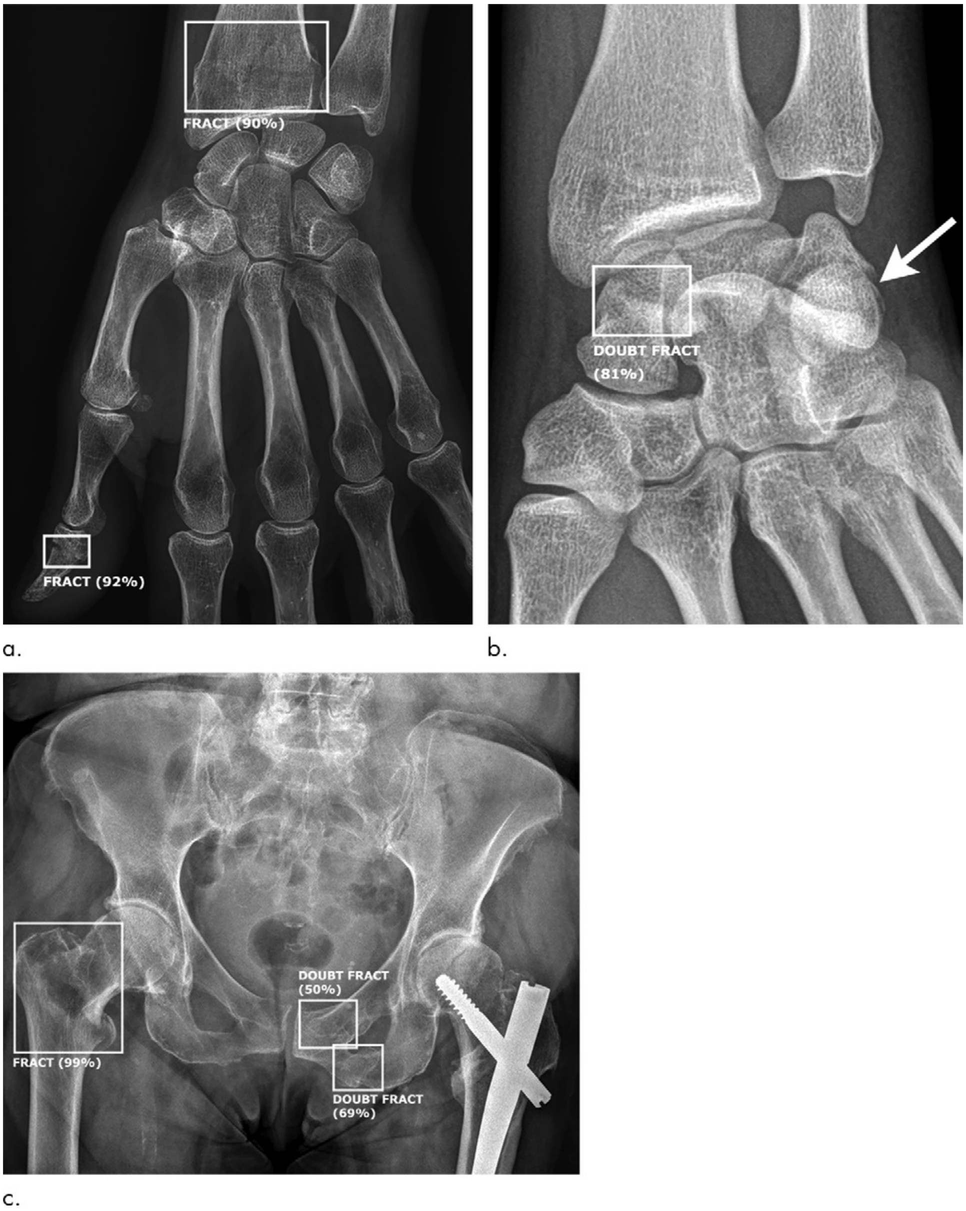

Figure 4: Radiographs show examples of multiple and/or severe fractures as well as human and artificial intelligence (AI) false-negative findings. AI system boxes are displayed as FRACT (definite fracture, confidence level >90%) and DOUBT FRACT (possible fracture, confidence level >50%) with confidence level, as explained in Materials and Methods. **(a)** Image shows right distal radius fracture combined with fracture of distal phalange of thumb. The latter was missed by one of six unaided readers. **(b)** Image shows fractures of scaphoid (box, found by AI but missed by five of six unaided readers) and triquetrum (arrow, missed by AI and five of six unaided readers) during perilunate dislocation. **(c)** Image shows fractures of right femoral neck and left superior and inferior ramus of pubis. The two latter fractures were missed by five of six unaided readers.

on plain radiographs could reduce error rates. In this study, we assessed the effect of an artificial intelligence (AI) aid on the diagnostic performance of six radiologists and six emergency physicians in detection of appendicular fractures on trauma radiographs. The AI aid improved the sensitivity of physicians from 70.8% to 79.4% (8.7% increase, P = .003 for superiority) and specificity from 89.5% to 93.6% (4.1% increase, P < .001 for noninferiority) and reduced the number of false-positive fractures per patient from 0.113 to 0.066 (41.9% decrease, P = .02), with no additional reading time (from 67.0 to 57.0 seconds, P = .12).

To our knowledge, this was the first study to assess the performance of AI-aided health professionals in seeking bone fractures on all appendicular radiographs. Published studies that have investigated deep learning approaches to bone fracture detection focused on single body parts, such as hips (26–29), wrists (19–21,25), shoulders (22), or ankles (23). Moreover, Blüthgen et al (21) validated their results on a monocentric external test set and other studies used internal test sets, whereas we gathered an external multicentric study sample including image acquisition systems not present in the development set. Unlike our study, all published studies considered the evaluation of fracture detection

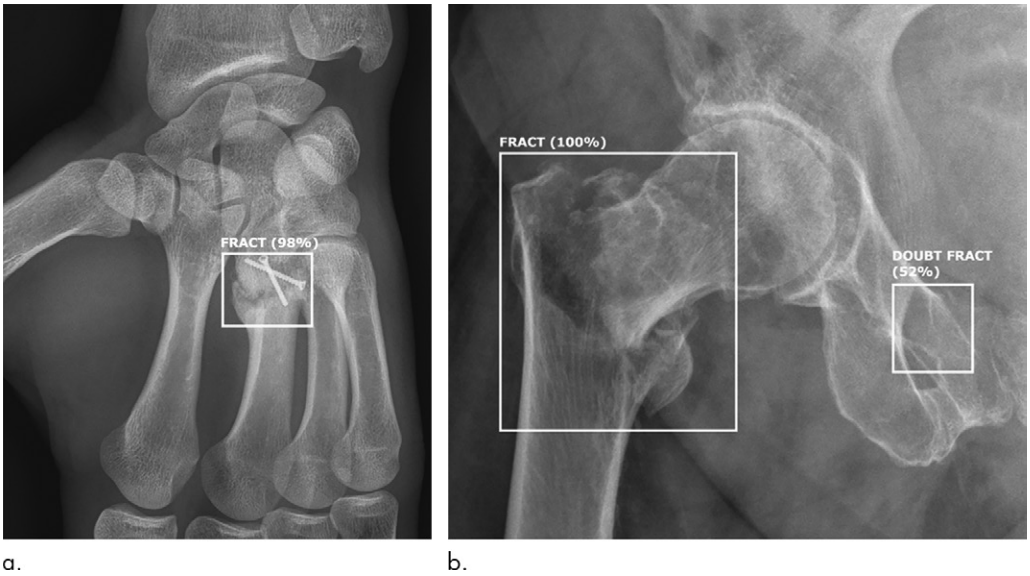

Figure 5: Radiographs show examples of false-positive findings with artificial intelligence (AI) system. AI system boxes are displayed as FRACT (definite fracture, confidence level >90%) and DOUBT FRACT (possible fracture, confidence level >50%) with confidence level, as explained in Materials and Methods. **(a)** Image shows left proximal second metacarpus healing fracture. The AI system indicated that it was a fracture; however, it is not a recent fracture and so was considered a false-positive finding. **(b)** Image shows true-positive AI finding of right femoral neck fracture (box on left of figure) and false-positive AI finding of fracture in pubic ramus (box on right of figure), possibly due to superposition of soft tissues

as a binary classification task, precluding the identification of multiple fractures on a single image, although it is one of the first sources of interpretation errors, known as satisfaction of search (32). Although most studies focused on the stand-alone AI performance, Lindsey et al (19) showed that the use of an AI aid could improve physician assistants' and emergency physicians' readings of wrist trauma radiographs. Moreover, whereas most previous studies used cropped or downsized images, our AI system handles full-resolution images with multiple radiographs per patient and can therefore be integrated into a viewer used in routine practice. Most published studies reported stand-alone algorithm AUCs higher than .90 or even .95, whereas we found AUCs of .94 for the stand-alone performance of our newer AI system when anatomic location was considered and .97 with use of a binary diagnosis without anatomic location, as did previous studies. We tried to obtain a setup that was as close as possible to real-word clinical conditions (whole appendicular body, multicentric external data set), with a strict definition of classification metrics, but we excluded obvious fractures from the study sample. Although excluding obvious fractures may have led to an underestimation of the overall diagnostic performance of unaided readers, it allowed us to show that an AI aid is helpful even in difficult but more clinically relevant conditions. Regardless, in our study, the stand-alone algorithm showed better performance than almost all readers, including radiologists, which has, to date, never been published.

Our study had several limitations. First, readers and the AI system were assessed on their ability to make decisions based on image analysis alone, without knowledge about the findings from the patients' physical examination or their medical history, creating a context bias (33). Clinical data can be crucial in making decisions (27,34); however, in our experience radiologists often lack relevant clinical data. Second, a Hawthorne effect may have affected the performances of readers, that is, a modification of their behavior in response to their awareness of being observed for the research project, leading, for instance, to a more thorough reading than in clinical practice. Similarly, cognitive biases related to the emergency setting could not be replicated in a retrospective study (35). Third, because examinations containing only obvious fractures were excluded, the sensitivity of unaided readers was probably underestimated. Fourth, the stratification of fractures, leading to an artificial 50% prevalence, made it impossible to calculate negative and positive predictive values and amplified the context bias. Fifth, a design in which all readers would have read the same images with and without AI might have yielded a higher statistical power. However, we chose a design that avoided reader-order bias (33), which is closer to a real-world setting.

In conclusion, we showed that a deep learning algorithm aided emergency physicians and radiologists in improving their diagnostic performance and boosting their time efficiency in the localization of all appendicular bone fractures on plain radiographs. The algorithm improved as updates were made, which bodes well for helping physicians cope with the increasing workload more effectively, and an evaluation in future prospective studies will be needed.

Author contributions: Guarantors of integrity of entire study, L.D., A.D., C.A., Z.Z.; 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, L.D., A.D., J.L., C.A., Z.Z., N.N., A.P., A. Feydy; clinical studies, L.D., A.D., C.A., Z.Z., L.L., N.E.R.; experimental studies, A.D., N.C., Z.Z., N.N., E.L.; statistical analysis, A.D., A.G., Z.Z., N.N., A.P.; and manuscript editing, L.D., A.D., A.G., J.L., Z.Z., N.N., A.P., L.L., A. Feydy

Disclosures of Conflicts of Interest: **L.D.** disclosed no relevant relationships. **A.D.** disclosed no relevant relationships. **A.G.** disclosed no relevant relationships. **J.L.** disclosed no relevant relationships. **C.A.** disclosed no relevant relationships. **N.C.** disclosed no relevant relationships. **Z.Z.** disclosed no relevant relationships. **N.N.** disclosed no relevant relationships. **E.L.** disclosed no relevant relationships. **A.P.** disclosed no relevant relationships. **A. Felter** disclosed no relevant relationships. **L.L.** Activities related to the present article: received a consulting fee from Gleamer for reading of gold standard examinations. Activities not related to the present article: received a consulting fee from Gleamer for reading of gold standard examinations. Other relationships: disclosed no relevant relationships. **L.L.** Activities related to the present article: received a consulting fee from Gleamer for reading of gold standard examinations. Activities not related to the present article: received a consulting fee from Gleamer for reading of gold standard examinations. Other relationships: disclosed no relevant relationships. **N.E.R.** Activities related to the present article: is co-founder and chief medical officer at Gleamer. Activities not related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. **A. Feydy** Activities related to the present article: received a consulting fee from Gleamer. Activities not related to the present article: received a consulting fee from Gleamer. Other relationships: disclosed no relevant relationships.

References

1. DiMaggio CJ, Avraham JB, Lee DC, Frangos SG, Wall SP. The Epidemiology of Emergency Department Trauma Discharges in the United States. Acad Emerg Med 2017;24(10):1244–1256.
2. Chen W, Lv H, Liu S, et al. National incidence of traumatic fractures in China: a retrospective survey of 512 187 individuals. Lancet Glob Health 2017;5(8): e807–e817.
3. Arasu VA, Abujudeh HH, Biddinger PD, et al. Diagnostic emergency imaging utilization at an academic trauma center from 1996 to 2012. J Am Coll Radiol 2015;12(5):467–474.
4. Willett JK. Imaging in trauma in limited-resource settings: a literature review. Afr J Emerg Med 2019;9(Suppl):S21–S27.
5. A national review of radiology reporting within the NHS in England. Care Quality Commission. https://www.cqc.org.uk/sites/default/files/20180718-radiology-reporting-review-report-final-for-web.pdf. Published July 2018. Accessed January 12, 2021.
6. Clinical radiology UK workforce census 2019 report. The Royal College of Radiologists. https://www.rcr.ac.uk/publication/clinical-radiology-uk-workforce-census-2019-report. Published 2019. Accessed January 12, 2021.
7. Tranovich MJ, Gooch CM, Dougherty JM. Radiograph interpretation discrepancies in a community hospital emergency department. West J Emerg Med 2019;20(4):626–632.
8. Scepi M, Rouffineau J, Faure JP, Richer JP, Van Der Marcq P. Discordant results in x-ray interpretations between ED physicians and radiologists. A prospective investigation of 30000 trauma patients. Am J Emerg Med 2005;23(7):918–920.
9. Communication of Radiograph Discrepancies between Radiology and Emergency Departments. Pennsylvania Patient Safety Advisory. http://patientsafety.pa.gov/ADVISORIES/Pages/201003_18.aspx. Accessed April 15, 2020.
10. Teixeira PGR, Inaba K, Salim A, et al. Preventable morbidity at a mature trauma center. Arch Surg 2009;144(6):536–541; discussion 541–542.
11. Guly HR. Diagnostic errors in an accident and emergency department. Emerg Med J 2001;18(4):263–269.
12. Whang JS, Baker SR, Patel R, Luk L, Castro A 3rd. The causes of medical malpractice suits against radiologists in the United States. Radiology 2013;266(2):548–554.
13. Fenton JJ, Taplin SH, Carney PA, et al. Influence of computer-aided detection on performance of screening mammography. N Engl J Med 2007;356(14):1399–1409.
14. Shaukat F, Raja G, Frangi AF. Computer-aided detection of lung nodules: a review. J Med Imaging 2019;6(2):020901.
15. Donnelley M, Knowles G, Hearn T. A CAD System for Long-Bone Segmentation and Fracture Detection. In: Elmoataz A, Lezoray O, Nouboud F, Mammass D, eds. Image and Signal Processing. ICISP 2008. Lecture Notes in Computer Science, vol 5099. Berlin, Germany: Springer, 2008; 153–162.
16. Cao Y, Wang H, Moradi M, Prasanna P, Syeda-Mahmood TF. Fracture detection in x-ray images through stacked random forests feature fusion. In: 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI), Brooklyn, NY, April 16–19, 2015. Piscataway, NJ: IEEE, 2015; 801–805.
17. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542(7639):115–118.
18. McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. Nature 2020;577(7788):89–94.
19. 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.
20. Kim DH, MacKinnon T. Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks. Clin Radiol 2018;73(5):439–445.
21. Blüthgen C, Becker AS, Vittoria de Martini I, Meier A, Martini K, Frauenfelder T. Detection and localization of distal radius fractures: deep learning system versus radiologists. Eur J Radiol 2020;126:108925.
22. Chung SW, Han SS, Lee JW, et al. Automated detection and classification of the proximal humerus fracture by using deep learning algorithm. Acta Orthop 2018;89(4):468–473.
23. Kitamura G, Chung CY, Moore BE 2nd. Ankle Fracture Detection Utilizing a Convolutional Neural Network Ensemble Implemented with a Small Sample, De Novo Training, and Multiview Incorporation. J Digit Imaging 2019;32(4):672–677.
24. Adams M, Chen W, Holcdorf D, McCusker MW, Howe PD, Gaillard F. Computer vs human: deep learning versus perceptual training for the detection of neck of femur fractures. J Med Imaging Radiat Oncol 2019;63(1):27–32.
25. Yahalomi E, Chernofsky M, Werman M. Detection of distal radius fractures trained by a small set of X-ray images and Faster R-CNN. http://arxiv.org/abs/1812.09025. Published December 21, 2018. Accessed September 26, 2019.
26. Gale W, Oakden-Rayner L, Carneiro G, Bradley AP, Palmer LJ. Detecting hip fractures with radiologist-level performance using deep neural networks. http://arxiv.org/abs/1711.06504. Published November 17, 2017. Accessed September 26, 2019.
27. Badgeley MA, Zech JR, Oakden-Rayner L, et al. Deep learning predicts hip fracture using confounding patient and healthcare variables. NPJ Digit Med 2019;2(1):31.
28. Jiménez-Sánchez A, Kazi A, Albarqouni S, et al. Towards an Interactive and Interpretable CAD System to Support Proximal Femur Fracture Classification. http://arxiv.org/abs/1902.01338. Published February 4, 2019. Accessed September 26, 2019
29. 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.
30. Bluemke DA, Moy L, Bredella MA, et al. Assessing radiology research on artificial intelligence: a brief guide for authors, reviewers, and readers-from the Radiology Editorial Board. Radiology 2020;294(3):487–489.
31. Wu Y, Kirillov A, Massa F, Lo WY, Girshick R. Detectron2. https://github.com/facebookresearch/detectron2. Published 2019. Accessed January 12, 2021.
32. Brady AP. Error and discrepancy in radiology: inevitable or avoidable? Insights Imaging 2017;8(1):171–182.
33. Gennaro G. The "perfect" reader study. Eur J Radiol 2018;103:139–146.
34. Loy CT, Irwig L. Accuracy of diagnostic tests read with and without clinical information: a systematic review. JAMA 2004;292(13):1602–1609.
35. Pines JM, Strong A. Cognitive Biases in Emergency Physicians: A Pilot Study. J Emerg Med 2019;57(2):168–172.