Artificial Intelligence for Detecting Acute Fractures in Patients Admitted to an Emergency Department: Real-Life Performance of Three Commercial Algorithms

Valérie Bousson, MD, PhD, Grégoire Attané, MD, Nicolas Benoist, MD, Laetitia Perronne, MD, Abdourahmane Diallo, PhD, Lama Hadid-Beurrier, PhD, Emmanuel Martin, PhD, Lounis Hamzi, MD, Arnaud Depil Duval, MD, Eric Revue, MD, Eric Vicaut, MD, PhD, Cécile Salvat, PhD

Rationale and Objectives: Interpreting radiographs in emergency settings is stressful and a burden for radiologists. The main objective was to assess the performance of three commercially available artificial intelligence (AI) algorithms for detecting acute peripheral fractures on radiographs in daily emergency practice.

Materials and Methods: Radiographs were collected from consecutive patients admitted for skeletal trauma at our emergency department over a period of 2 months. Three AI algorithms—SmartUrgence, Rayvolve, and BoneView—were used to analyze 13 body regions. Four musculoskeletal radiologists determined the ground truth from radiographs. The diagnostic performance of the three AI algorithms was calculated at the level of the radiography set. Accuracies, sensitivities, and specificities for each algorithm and two-by-two comparisons between algorithms were obtained. Analyses were performed for the whole population and for subgroups of interest (sex, age, body region).

Results: A total of 1210 patients were included (mean age 41.3 ± 18.5 years; 742 [61.3%] men), corresponding to 1500 radiography sets. The fracture prevalence among the radiography sets was 23.7% (356/1500). Accuracy was 90.1%, 71.0%, and 88.8% for SmartUrgence, Rayvolve, and BoneView, respectively; sensitivity 90.2%, 92.6%, and 91.3%, with specificity 92.5%, 70.4%, and 90.5%. Accuracy and specificity were significantly higher for SmartUrgence and BoneView than Rayvolve for the whole population ($P < .0001$) and for subgroups. The three algorithms did not differ in sensitivity ($P = .27$). For SmartUrgence, subgroups did not significantly differ in accuracy, specificity, or sensitivity. For Rayvolve, accuracy and specificity were significantly higher with age 27-36 than ≥53 years ($P = .0029$ and $P = .0019$). Specificity was higher for the subgroup knee than foot ($P = .0149$). For BoneView, accuracy was significantly higher for the subgroups knee than foot ($P = .0006$) and knee than wrist/hand ($P = .0228$). Specificity was significantly higher for the subgroups knee than foot ($P = .0003$) and ankle than foot ($P = .0195$).

Conclusion: The performance of AI detection of acute peripheral fractures in daily radiological practice in an emergency department was good to high and was related to the AI algorithm, patient age, and body region examined.

Key Words: Artificial intelligence; Deep learning; Medical imaging; Musculoskeletal imaging; Fracture.

Abbreviations: AI artificial intelligence, **BCa bootstrap** bias-corrected and accelerated bootstrap, **GEE** generalized estimating equation, **PACS** Picture Archiving and Communication System (PACS), **Q1-Q3** interquartile range, **SD** standard deviation (SD)

Acad Radiol 2023; 30:2118–2139

From the Radiology Department, Lariboisière's Hospital, AP-HP.Nord-Université de Paris, 2 rue Ambroise Paré, 75010, Paris, France (V.B., G.A., N.B., L.P., L.H.); Clinical Research Department, Lariboisière's Hospital, AP-HP.Nord-Université de Paris, Paris, France (A.D., E.V.); Medical Physics Department, Lariboisière's Hospital, AP-HP.Nord-Université de Paris, Paris, France (L.H.-B., C.S.); Information Technology Department, Lariboisière's Hospital, AP-HP.Nord-Université de Paris, Paris, France (E.M.); Emergency Department, Lariboisière's Hospital, AP-HP.Nord-Université de Paris, Paris, France (A.D.D., E.R.); Emergency Department, Saint-Joseph's Hospital, Paris, France (A.D.D.). Received April 11, 2023; revised June 8, 2023; accepted June 20, 2023. **Address correspondence to**: V.B. e-mail: valerie.bousson@aphp.fr

https://doi.org/10.1016/j.acra.2023.06.016

INTRODUCTION

In multiple medical domains, including medical image analysis, artificial intelligence (AI) has proven its ability to augment physician performance (1–5). Use of AI is especially interesting for radiographic fracture detection. Indeed, interpreting radiographs for detecting fractures in emergency settings is a huge task and carries significant risk of diagnostic errors, some responsible for severe consequences (6,7). Many studies have reported the high diagnostic performance of AI algorithms in this task, with a pooled sensitivity of 91% and specificity of 94% in a recent meta-analysis of 42 peer-reviewed publications (8). Initially AI algorithms were trained on specific body parts, and the ability to recognize a fracture was specific to that body part (8).

AI algorithms able to detect fractures of the whole peripheral skeleton are commercially available to assist radiologists and emergency physicians in real life. A retrospective study using such an algorithm (BoneView) demonstrated in a case-control series of 600 patients that AI assistance improved the sensitivity of readers in detecting fractures by 8.7% and specificity by 4.1% (9). Similar results were obtained in another retrospective study using the same algorithm with a data set of 480 examinations (10). A different commercially available algorithm, SmartUrgence, was recently tested in a multireader diagnostic accuracy study (300 musculoskeletal, chest, and abdominal radiographs) to determine whether it could pass the radiographic reporting component of the Fellowship of the Royal Collège of Radiologists and be compared to readings by certified radiologists (11). Radiologists achieved an average accuracy of 84.8% and SmartUrgence an accuracy of 79.5%. A third commercially available algorithm, Rayvolve, evaluated in a large real-life cohort of children (*n* = 2549, presenting as routine to the emergency room), was found very reliable for fracture detection with an accuracy > 90% (12). However, the current performance of the use of commercial AI algorithms in detecting fracture in a clinical emergency practice in adults remains incompletely evaluated because of the design of these studies with their selection of cases and good-quality radiographs (9–11), a fracture prevalence set at 50% (9–11), or a specific study population (12).

Therefore, our main objective was to assess the performance of three commercially available AI algorithms in detecting acute fractures (9–12), with the inclusion of non selected radiographs from consecutive patients admitted to an emergency department for acute skeletal trauma and unknown prevalence of fractures. We examined the performance in the whole population and according to sex, age, or body region. Our secondary objective was to compare the performance of the three AI algorithms with each other.

MATERIALS AND METHODS

The study was conducted in the radiology department of Lariboisière hospital, Paris (Assistance Publique-Hôpitaux de Paris-Université Paris Cité), in collaboration with the emergency department. The main steps consisted of (1) implementing the three AI algorithms for automatic detection and location of peripheral fractures and detecting fractures from March to May 2021; (2) collecting all consecutive multiview radiography sets and AI reports (AI prediction) performed from June 1 to end of July 2021; (3) establishing the radiologist diagnosis and rating the algorithms from October to November 2021; and (4) submitting the files to our statistical department for analyses in December 2021 (Fig 1). The three AI algorithms were SmartUrgence (v1.7, Milvue, France, CE Certificate of Conformity), Rayvolve (AZmed, France, CE Certificate of Conformity, US Food and Drug Administration [FDA] approval), and BoneView (v1.0.2, Gleamer, France, CE Certificate of Conformity, FDA approval). Ethical approval for the study was granted by the institutional research ethics committee (CRM-2110-210) that waived informed consent because of the retrospective nature of the analysis of radiographs, AI reports, and epidemiological data.

Implementation of AI Algorithms

Milvue, AZmed, and Gleamer provided the latest edition of their AI solution, and no change in the algorithm was allowed during the period of evaluation. None of the algorithms was trained with radiographs from our center before implementation.

This step consisted of solving technical issues related to the transfer of digital radiographs from the radiography emergency room for the application of each of the three AI algorithms and transfer of algorithm reports to our institutional Picture Archiving and Communication System (PACS) (Carestream, v12.1.6.0117). The most operational solution was for radiographers to manually transfer all radiographs to the PACS and each of the three AI algorithms. After analysis of each radiograph, the three AI algorithms automatically sent to the PACS an annotated report for each processed radiograph within less than 1 minute. In the reports, AI-detected fractures were surrounded by a box in a continuous white line (Fig 2). Doubtful fractures for SmartUrgence and BoneView were surrounded by a white dotted box. The doubtful fracture item was not present for Rayvolve. Images were displayed on the PACS in the following order: multiview radiography set for one or several body regions, then the set of AI reports for each AI algorithm for each radiograph.

Study Population (Table 1)

Radiography sets were collected from all consecutive patients who underwent radiography after admission for peripheral skeletal trauma to our emergency department. Lariboisière hospital, Paris (Assistance Publique-Hôpitaux de Paris-Université Paris Cité), emergency department is a general emergency center. According to an estimated prevalence of peripheral fractures of 15%-20% among patients who underwent radiography, the size of study populations, and the AI algorithm performance from previous studies (9–12), we aimed to collect 1500 radiography sets, with one set corresponding to one body region. The inclusion criteria were (1) age 15 years or older; (2) admission to the emergency department for recent

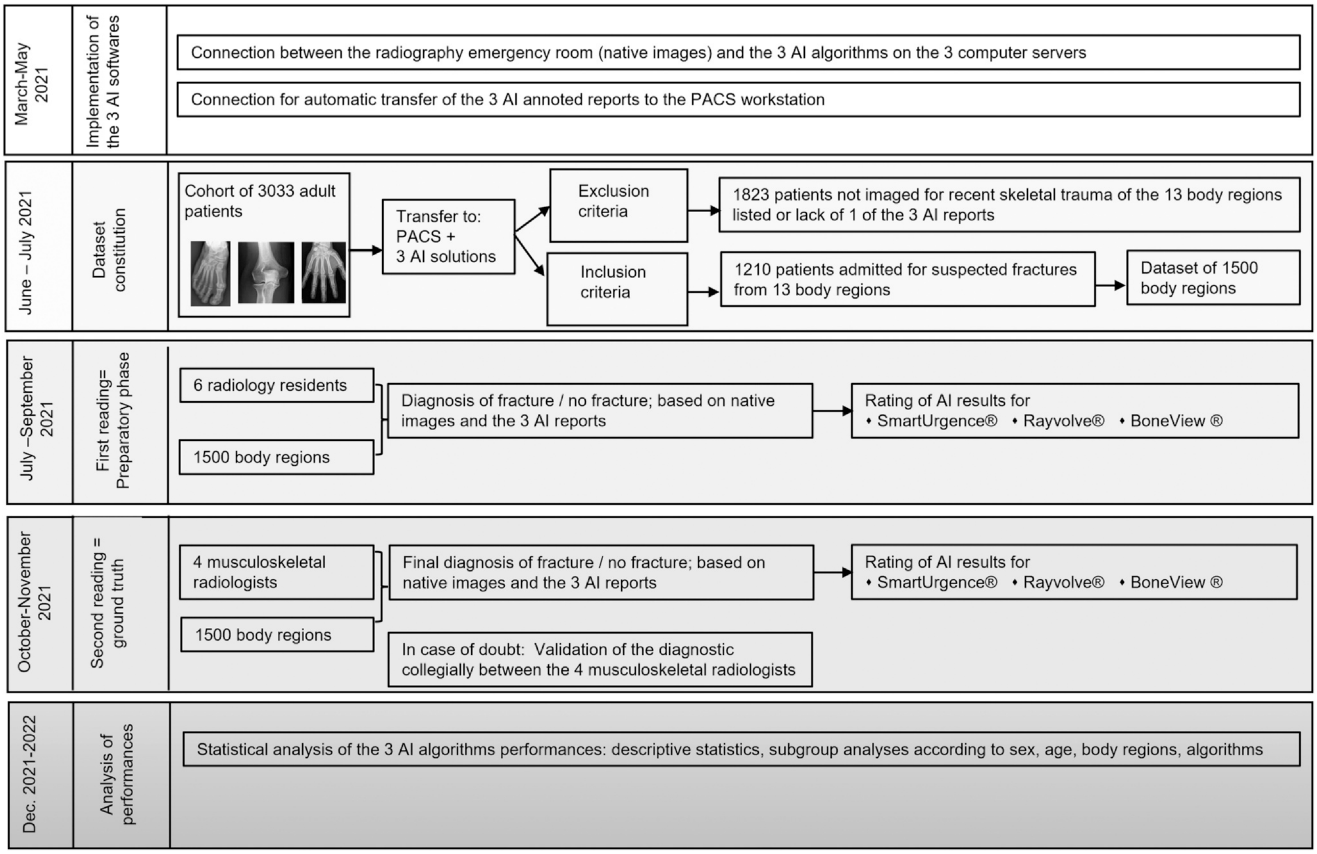
March-May 2021
Implementation of the 3 AI softwares
Connection between the radiography emergency room (native images) and the 3 AI algorithms on the 3 computer servers
Connection for automatic transfer of the 3 AI annoted reports to the PACS workstation
June – July 2021
Dataset constitution
Cohort of 3033 adult patients
Transfer to: PACS + 3 AI solutions
Exclusion criteria
1823 patients not imaged for recent skeletal trauma of the 13 body regions listed or lack of 1 of the 3 AI reports
Inclusion criteria
1210 patients admitted for suspected fractures from 13 body regions
Dataset of 1500 body regions
July –September 2021
First reading= Preparatory phase
6 radiology residents
1500 body regions
Diagnosis of fracture / no fracture; based on native images and the 3 AI reports
Rating of AI results for • SmartUrgence® • Rayvolve® • BoneView ®
October-November 2021
Second reading = ground truth
4 musculoskeletal radiologists
1500 body regions
Final diagnosis of fracture / no fracture; based on native images and the 3 AI reports
Rating of AI results for • SmartUrgence® • Rayvolve® • BoneView ®
In case of doubt: Validation of the diagnostic collegially between the 4 musculoskeletal radiologists
Dec. 2021-2022
Analysis of performances
Statistical analysis of the 3 AI algorithms performances: descriptive statistics, subgroup analyses according to sex, age, body regions, algorithms

Figure 1. Flowchart of the study. AI, artificial intelligence; PACS, Picture Archiving and Communication System.

skeletal trauma; (3) one or several body regions examined by radiography from the following 13 body regions (acceptable for the three AI algorithms): clavicle, shoulder, humerus shaft, elbow, radius/ulna shaft, wrist/hand, finger, pelvis/hip, femur shaft, knee, tibia/fibula shaft, ankle, and foot; and (4) AI reports available in the PACS for the three algorithms for the whole set of radiographs for each patient. We did not exclude sets on the basis of a poor-quality radiograph or the presence of a cast or metallic implant.

Post-traumatic radiographs were acquired on a Clisis Exel DRF digital radiograph instrument (Primax-GMM Group, Seriate). Immediately after acquisition, radiographs were submitted for analysis by the three AI algorithms. Radiographs and AI reports were available from the PACS for emergency physicians. No report was provided by radiologists. Radiography sets and AI reports were consecutively collected from June 1 to the end of July. Analyses of radiographs and AI reports by radiologists were retrospective.

Methodology for Algorithm Evaluation (Fig. 1and 2)

Preparatory Phase

The preparatory phase consisted of the data preparation by radiology residents. All radiography sets acquired from 1 of 6 days were given to each of six radiology residents (4 years of residency) for analysis until complete consecutive analysis of 250 multiview radiography sets per resident (= 250 body regions). The residents recorded in a Microsoft Excel spreadsheet the day of examination, patient-anonymized identity, sex, age, and body region. Considering the whole set of available images that were the radiographs and AI reports, the residents proposed for each radiography set the diagnosis of fracture or no fracture and rated the AI algorithm reports. This phase was performed to facilitate the subsequent analysis work by the tenured radiologists and, in our department with an academic mission, to play an educational role for residents on the potential of AI in radiology (13). To rate algorithm reports, four decisions were assumed: (1) correct AI identification of a fracture was defined as a box centered on the fracture; (2) for a harmonization issue between algorithms, doubtful AI fractures (SmartUrgence and BoneView) were systematically classified as AI fracture; (3) AI identification of an old fracture (well-defined sclerotic borders) was considered a false-positive identification; and (4) the basic unit was the body region (= radiography set). For each body region, the AI report could identify zero, one, or several fractures. For that body region, the radiologist's diagnosis was zero, one, or several fractures. Therefore, the reader assigned one of five categories (A-E) to the AI report of each body region: (A) true-positive and/or true-negative identifications; (B) true-positive plus false-positive identifications; (C) false-positive identification; (D) true-positive plus false-negative identifications; and (E) false-negative identification (s). Category A is the perfect category. In categories B and C, the

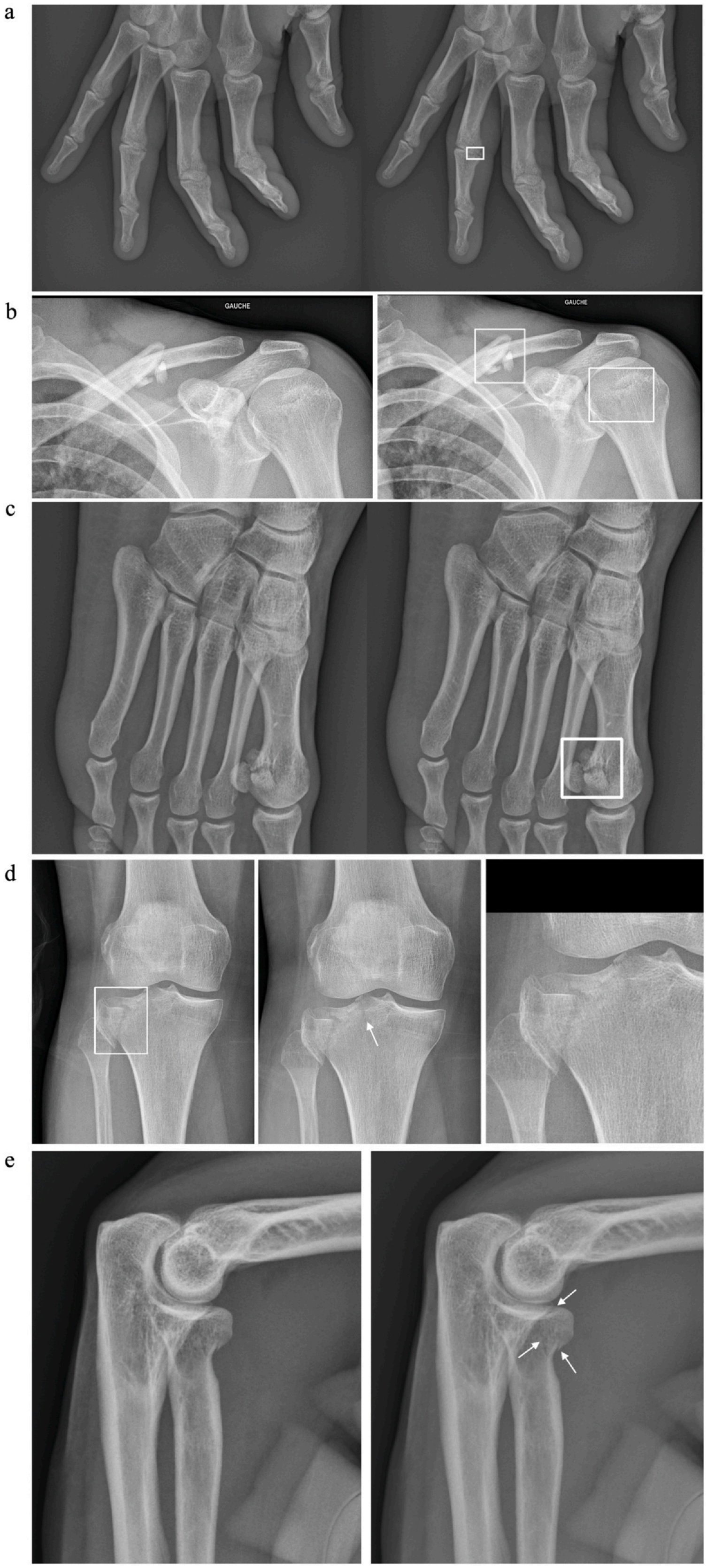

Figure 2. Cases illustrating the rating in five categories. (**a**) Fracture of the proximal phalanx correctly identified by Artificial Intellignece (AI). No error. (**b**) Fracture of the clavicle correctly identified by AI. False-positive identification of a fracture of the proximal humerus. (**c**) False-positive identification of a fracture (bipartite sesamoid) by IA. The radiograph was obtained to search for a fracture at the base of the fifth metatarsal. (**d**) Fracture of the lateral tibial plateau correctly identified by AI. Missed fracture of the spinal process (*white arrow*). (**e**) The fracture of the radial head (*white arrow*) was missed by IA. AI, artificial intelligence;

algorithm provides one or more erroneous identifications of fracture. In categories D and E, one or more fractures are missed.

Final Diagnosis and Rating of Algorithms Reports

Two months later, four musculoskeletal radiologists, three fellows (Grégoire Attané, Nicolas Benoist, and Laetitia Perronne, with 1-2 years of fellowship) and one senior radiologist (Valérie Bousson, with 20 years of experience), reanalyzed the radiography sets. Indeed, the same data spreadsheet was split into four data sets. Musculoskeletal radiologists provided a final diagnosis and the algorithms' rating based on the analysis of radiographs, AI reports, information related to the follow-up if available, and resident categories. In case of doubt, the final diagnosis was established collegially between the four radiologists. The diagnosis of radiologists was used as the ground truth, and their categories were used for statistical analyses to evaluate the performance of the algorithms.

Statistical Analyses

Continuous variables are reported as mean with standard deviation (SD) and median with interquartile range (Q1-Q3). Categorical variables are reported as number (percentage) and were compared by χ^2 test or Fisher's exact test.

The diagnostic performance of AI algorithms was calculated at the level of the radiography set (see A/B/C/D/E categories defined previously). For each AI algorithm, statistical distributions of these categories were compared by subgroups of interest (ie, sex, age, and body region) by χ^2 test or Fisher's exact test, if appropriate.

Two levels of performance were considered to compare the usefulness of the IA algorithms:

(1) *The accuracy of each algorithm*, defined as its ability to make the exact diagnosis defined as the ratio of category A to the total number of radiography sets analyzed.

(2) *The clinically oriented performance of each algorithm,* defined as its ability to detect at least one existing fracture among radiography sets with at least one fracture (even if some additional incorrect fracture diagnoses were made) (ie, corresponding to the ratio of category A + B + D to the number of radiography sets with at least one fracture). Such a definition can be considered as *sensitivity per radiography set* and is close to the sensitivity proposed in (9). *The specificity per radiography set* is defined as the proportion of radiography sets in which no fracture was detected among sets with no fracture, a definition that is also similar to the specificity used in (9).

For all calculations, we accounted for the inter-radiography set correlation within the same patient by considering patient as a cluster. Thus, generalized estimating equations (GEEs) were used to calculate proportions, and the cluster bootstrap bias-corrected and accelerated (BCa) bootstrap method was used for 95% confidence interval (95% CI) calculation. We also used GEE logistic regressions to compare the three AI algorithms globally at a 5% two-sided significance level. Two-by-two comparisons between algorithms also involved the same methods, but the α value was adjusted for multiplicity with the simulation-based method (Edwards et al., proc glimmix SAS, adjust = simulate).

These analyses were performed on the total population and prespecified subgroups (sex, age, body region). Subgroup analyses should be considered exploratory.

All analyses were performed with SAS 9.4 (SAS Institute Inc., Cary, NC, USA).

RESULTS

Study Population

The constitution of the study population and radiography sets is described in Figure 1. Descriptive statistics of the study population and radiography sets are provided in Table 1.

We obtained 1500 radiography sets for 1210 consecutive patients; the mean age of the patients was 41.3 ± 18.5 years; 61.3% (742/1210) were men. Most patients (82.0%, 992/1210) had only one body region examined; 27.7% (335/1210) had at least one fracture.

The most frequently examined body regions were the wrist/hand (20.9% [314/1500]), ankle (15.5% [232/1500]), knee (13.1% [197/1500]), and foot (12.4% [186/1500]). There were 356 (23.7%) fractured regions among the 1500 radiography sets; 222 (62.4%) fractures were located in the upper extremities and shoulder girdle and 134 (37.6%) in the lower extremities and pelvic girdle. Among the radiography sets, four regions had a fracture rate > 30%: humerus (35.7%, 5/14), elbow (31.2%, 38/119), wrist/hand (30.8%, 97/314), and foot (30.1%, 56/186). The knee was the less fractured region (6.6%, 13/197). Eleven radiography sets had been read in consensus (0.7%, 11/1500, no fracture in three sets, one or two fractures in eight sets); body regions were the clavicle ($n = 1$), elbow ($n = 2$), wrist/hand ($n = 4$), and foot ($n = 4$).

Descriptive Statistics of AI Algorithm Results

Categories obtained by the three AI algorithms are described for the 1500 radiography sets in Table A1 and by subgroups (sex, four subgroups of age, and four most frequently examined body regions) in Table A2.

For the 1500 radiography sets, SmartUrgence, Rayvolve, and BoneView provided the perfect category (category A: true-positive and/or true-negative identifications) in 90.4% (1356/1500), 71.3% (1070/1500), and 89.1% (1336/1500) of sets, respectively (Table A1); false-positive identifications in 7%, 26.5%, and 8.7% of sets; and false-negative identifications in 2.7%, 2.2%, and 2.2% of sets. A total of 57 radiography sets had a missed fracture, 45 had a single fracture, and 12 multiple fractures. Among the 45 radiography sets with a single fracture (45/57), in 9 (9/57), the three algorithms provided a report with a false-negative identification (no particular body region); for 16 sets, two

TABLE 1. Descriptive Statistics: (a) Demographic Characteristics ($n = 1210$) and (b) Descriptive Statistics of the Radiography Sets ($n = 1500$)

	Overall Population (Patients) $N = 1210$
(a)	
Age, y	
n (miss.)	1210 (0)
Mean ± SD	41.3 ± 18.5
Median (Q1; Q3)	37 (27; 52)
Min, max	15, 104
Sex, no. (%)	
Male	742 (61.3%)
Female	468 (38.7%)
All	1210 (100.0%)
Body region examined, no. (%)	
6	2 (0.2%)
5	3 (0.2%)
4	7 (0.6%)
3	41 (3.4%)
2	165 (13.6%)
1	992 (82.0%)
All	1210 (100.0%)
Fractures, no. (%)	
2	21 (1.7%)
1	314 (26.0%)
0	875 (72.3%)
All	1210 (100.0%)
	Radiography sets $N = 1500$
(b)	
Radiography set, body region, no. (%)	
Clavicule	59 (3.9%)
Shoulder	149 (9.9%)
Humerus, diaphysis	14 (0.9%)
Elbow	119 (7.9%)
Radius/ulna, diaphysis	20 (1.3%)
Wrist/hand	314 (20.9%)
Finger	103 (6.9%)
Pelvis/hip	49 (3.3%)
Femur, diaphysis	17 (1.1%)
Knee	197 (13.1%)
Tibia/fibula, diaphysis	41 (2.7%)
Ankle	232 (15.5%)
Foot	186 (12.4%)
All	1500 (100.0%)
Final diagnosis of senior radiologist, no. (%)	356 (23.7%)
Fracture	1144 (76.3%)
No fracture	1500 (100.0%)
All	
Fracture, body region, no. (%)	
Clavicule	13 (3.6%)
Shoulder	39 (11%)
Humerus, diaphysis	5 (1.4%)
Elbow	38 (10.7%)
Radius/ulna, diaphysis	4 (1.1%)
Wrist/hand	97 (27.2%)
Finger	26 (7.3%)
Pelvis/hip	13 (3.7%)
Femur, diaphysis	3 (0.8%)
Knee	13 (3.7%)
Tibia/fibula, diaphysis	7 (2%)
Ankle	42 (11.8%)
Foot	56 (15.7%)
All	356 (100%)

In (a): data are number (%) of patients.
In (b): data are number (%) of radiography sets.
Q, quartile; SD, standard deviation.

algorithms provided a false-negative identification; and for 20 sets, only one algorithm provided a false-negative identification. For the 12 (12/57) sets with multiple fractures, except for one time, the three algorithms always correctly identified at least one fracture.

For the 11 radiography sets read in consensus, for 5 sets the three AI algorithms provided similar reports (correct diagnosis of no fracture, $n = 2$; correct diagnosis of fracture, $n = 2$; false-positive identification, $n = 1$). For six sets, the AI algorithms provided different reports (categories for SmartUrgence, Rayvolve, and BoneView were set 1: A-A-D; set 2: E-A-A; set 3: B-B-A; set 4: E-B-B; set 5: E-A-A; and set 6: A-A-E).

Categories did not significantly differ by sex for SmartUrgence ($P = .34$), Rayvolve ($P = .62$), or BoneView ($P = .48$) (Table A2a). For Rayvolve, categories significantly differed by age class ($P < .01$) (Table A2b). Categories significantly differed by body region for Rayvolve and BoneView ($P = .0002$ and $P < .0001$, respectively) but not for SmartUrgence ($P = .1761$) (Table A2c).

Performance of Each AI Algorithm for All Radiography Sets and Two-by-Two Comparisons of the Performance of the Algorithms

Accuracy, sensitivity, and specificity for each of the three AI algorithms for all radiography sets are reported in Tables 2-5. For all radiography sets, accuracy was 90.1%, 71.0%, and 88.8% for SmartUrgence, Rayvolve, and BoneView, respectively; sensitivity 90.2%, 92.6%, and 91.3%; and specificity 92.5%, 70.4%, and 90.5%.

Two-by-two comparisons of the algorithms are reported in Tables 2–5**.** Accuracy and specificity were significantly higher for SmartUrgence and BoneView than Rayvolve ($P < .0001$ and $P < .0001$ for SmartUrgence vs. Rayvolve; $P < .0001$ and $P < .0001$ for BoneView vs. Rayvolve). Sensitivity did not significantly differ between the three algorithms ($P = .2650$). Accuracy, sensitivity, and specificity did not significantly differ between SmartUrgence and BoneView.

TABLE 2. Diagnostic Performance, Accuracy (1500 Radiography Sets; Subgroup Sex; Subgroup Age; Subgroup Body Region)

	Diagnostic Performance: Accuracy						
	AI Algorithms			Mean Difference			
	SmartUrgence	Rayvolve	BoneView	SmartUrgence- vs. Rayvolve	BoneView vs. Rayvolve	SmartUrgence vs. BoneView	*P* Value*
Accuracy							**<.0001**
All radiography sets	90.1% (88.6%; 91.5%)[a]	71.0% (68.8%; 72.9%)[a]	88.8% (87.4%; 90.2%)[a]	19.2% (16.3%; 22.0%) ***P*** ** **< .0001**	17.9% (15.0%; 20.7%) ***P*** < .0001**	1.3% (−0.8%; 3.4%) *P*** = .32	
Gender							**.7642**~
Female	90.8% (88.5%; 92.8%)[a]	72.1% (68.8%; 75.1%)[a]	90.3% (87.8%; 92.4%)[a]	18.7% (13.3%; 24.0%) ***P*** < .0001**	18.1% (12.7%; 23.6%) ***P*** < .0001**	0.5% (−3.6%; 4.6%) *P*** = .9989	
Male	89.8% (87.9%; 91.7%)[a]	70.2% (67.4%; 73.1%)[a]	87.9% (86.0%; 89.7%)[a]	19.5% (15.1%; 24.0%) ***P*** < .0001**	17.7% (13.1%; 22.2%) ***P*** < .0001**	1.9% (−1.5%; 5.2%) *P*** = .5863	
Male vs. female	1.0% (−5.6%; 3.5%) *P*** = .9871	−1.9% (−8.8%; 5.0%) *P*** = .9690	−2.4% (−7.1%; 2.3%) *P*** = .6924				
Age class							**.0244~**
Q1: ≤26	90.0% (87.0%; 92.8%)[a]	72.4% (67.9%; 76.6%)[a]	89.8% (86.7%; 92.5%)[a]	17.6% (9.7%; 25.5%) ***P*** < .0001**	17.3% (9.6%; 25.0%) ***P*** < .0001**	0.3% (−4.3%; 4.8%) *P*** = 1.0000	
Q2: 27-36	90.5% (87.5%; 93.2%)[a]	77.1% (73.1%; 0.9%)[a]	89.7% (87.2%; 2.5%)[a]	13.4% (6.3%; 20.6%) ***P*** < .0001**	12.7% (4.5%; 20.8%) ***P*** < .0001**	0.8% (−5.4%; 7.0%) *P*** = 1.0000	
Q3: 37-52	92.1% (89.5%; 94.5%)[a]	70.7% (66.4%; 74.7%)[a]	88.5% (85.4%; 1.3%)[a]	21.4% (13.3%; 29.4%) ***P*** < .0001**	17.8% (10.2%; 25.4-%) ***P*** < .0001**	3.6% (−2.5%; 9.7%) *P*** = .7263	
Q4: ≥53	87.8% (84.6%; 90.7%)[a]	63.6% (59.0%; 7.7%)[a]	87.3% (83.9%; 90.0%)[a]	24.2% (16.2%; 32.2%) ***P*** < .0001**	23.7% (15.4%; 32.0-%) ***P*** < .0001**	0.5% (−5.9%; 6.9%) *P*** = 1.0000	
Q2: 27-36 vs. Q1: ≤26	0.5% (−7.0%; 7.9%)	4.6% (−6.1%; 15.3%)	0.0% (−7.4%;7.4%)				

TABLE 2 (Continued)

	Diagnostic Performance: Accuracy						
	AI Algorithms			Mean Difference			
	SmartUrgence	Rayvolve	BoneView	SmartUrgence-vs. Rayvolve	BoneView vs. Rayvolve	SmartUrgence vs. BoneView	*P* Value[*]
	*P*** = 1.0000	*P*** = .9563	*P*** = 1.0000				
Q2: 27-36 vs. Q3: 37-52	–1.6% (–8.5%; 5.3%) *P*** = .9998	6.3% (–4.3%; 16.9%) *P*** = .7147	1.2% (–6.1%; 8.6%) *P*** = 1.0000				
Q2: 27-36 vs. Q4: ≥53	2.7% (–4.8%; 10.2%) *P*** = .9892	13.5% (2.7%; 24.2%) ***P*** = .0029**	2.5% (–5.1%; 10.0%) *P*** = .9953				
Q3: 37-52 vs. Q1: ≤26	2.1% (–5.0%; 9.1%) *P*** = .9984	–1.7% (–12.8%; 9.4%) *P*** = 1.0000	–1.2% (–8.9%; 6.4%) *P*** = 1.0000				
Q3: 37-52 vs. Q4: ≥53	4.3% (–2.8%; 11.4%) *P*** = .7000	7.1% (–4.0%; 18.3%) *P*** = .6210	1.3% (–6.5%; 9.0%) *P*** = 1.0000				
Q4: ≥53 vs. Q1: ≤26	–2.2% (–9.9%; 5.4%) *P*** = .9983	–8.8% (–20.1%; 2.4%) *P*** = .2947	–2.5% (–10.3%; 5.3%) *P*** = .9962				
Body region							**.3269~**
Foot	87.0% (82.9%; 92.5%)[a]	66.3% (59.9%; 74.1%)[a]	82.1% (77.0%; 88.1%)[a]	20.7% (8.9%; 32.4%) ***P***** < .0001**	15.8% (4.7%; 26.8%) ***P*** = .0004**	4.9% (–5.7%; 15.5%) *P*** = .9203	
Knee	94.3% (90.9%; 97.7%)[a]	80.6% (75.2%; 86.6%)[a]	96.4% (93.3%; 98.7%)[a]	13.8% (4.7%; 22.8%) ***P*** = .0002**	15.8% (6.6%; 25.0%) ***P***** < .0001**	-2.0% (–8.2%; 4.1%) *P*** = .9939	
Ankle	90.2% (87.5%; 94.4%)[a]	70.0% (65.6%; 77.6%)[a]	91.9% (89.1%; 95.8%)[a]	20.2% (10.0%; 30.4%) ***P*** = <.0001**	21.9% (12.2%; 31.7-%) ***P***** < .0001**	–1.7% (–9.4%; 6.0%) *P*** = .9999	
Wrist/hand	88.4% (85.2%; 92.2%)[a]	72.8% (68.1%; 77.9%)[a]	88.4% (85.2%; 91.6%)[a]	15.7% (7.4%; 23.9%) ***P***** < .0001**	15.7% (7.1%; 24.2%) ***P***** < .0001**	0.0% (–6.6%; 6.6%) *P*** = 1.0000	
Knee vs. foot	7.3% (–2.4%; 17.1%) *P*** = .3423	14.3% (–0.9%; 29.4%) *P*** = .0841	14.3% (3.8%; 24.8%) ***P*** = .0006**				

TABLE 2 (Continued)

	Diagnostic Performance: Accuracy						
	AI Algorithms			Mean Difference			
	SmartUrgence	Rayvolve	BoneView	SmartUrgence-vs. Rayvolve	BoneView vs. Rayvolve	SmartUrgence vs. BoneView	*P* Value*
Knee vs. ankle	4.2% (−4.6%; 13.0%) *P*** = .9114	10.6% (−4.3%; 25.5%) *P*** = .4239	4.5% (−3.7%; 12.6%) *P*** = .7898				
	Knee vs. wrist/hand	5.9% (−2.1%; 14.0%)	7.8% (−5.1%; 20.7%)	8.0% (0.6%; 15.4%)			
*P*** = .6695	***P*** = .0228**						
Ankle vs. foot	3.2% (−5.9%; 12.2%) *P*** = .9900	3.6% (−11.3%; 18.6%) *P*** = .9996	9.8% (−1.1%; 20.7%) *P*** = .1236				
Ankle vs. wrist/hand	1.8% (−7.0%; 10.5%) *P*** = .9999	−2.8% (−16.2%; 10.6%) *P*** = .9999	3.5% (−4.9%; 11.9%) *P*** = .9637				
Wrist/hand vs. foot	1.4% (−8.5%; 11.4%) *P*** = 1.0000	6.4% (−8.4%; 21.2%) *P*** = .9487	6.3% (−4.8%; 17.5%) *P*** = .7537				

[a] 95% confidence intervals using bootstrap bias-corrected and accelerated method.

AI, artificial intelligence.

* Type III *P* values.

** *P* value adjustment for multiple comparisons using the simulation-based method.

~ *P* values for interaction between AI algorithms and subgroups.

TABLE 3. Diagnostic Performance, Sensitivity (1500 Radiography Sets; Subgroup Sex, Subgroup Age; Subgroup Body Region)

	Diagnostic Performance: Sensitivity						
	AI Algorithms			Mean Difference			
	SmartUrgence	Rayvolve	BoneView	SmartUrgencevs. Rayvolve	BoneView vs. Rayvolve	SmartUrgencevs. BoneView	P Value*
Sensitivity							.2650
All radiography sets	90.2% (87.2%; 92.8%)[a]	92.6% (90.1%; 94.6%)[a]	91.3% (88.2%; 93.6%)[a]	−2.5% (−6.0%; 1.1%) P** = .2343	−1.3% (−5.1%; 2.5%) P** = .6978	−1.1% (−4.9%; 2.6%) P** = .7554	
Gender							.8107~
Female	90.8% (86.9%; 94.4%)[a]	93.3% (89.8%; 96.4%)[a]	93.3% (89.9%; 96.2%)[a]	−2.5% (−8.5%; 3.5%) P** = .8467	0.0% (−6.0%; 6.0%) P** = 1.0000	−2.5% (−9.0%; 4.0%) P** = .8847	
Male	90.5% (86.5%; 93.7%)[a]	92.5% (89.2%; 95.6%)[a]	91.0% (87.0%; 94.3%)[a]	−2.1% (−7.1%; 2.9%) P** = .8386	−1.6% (−7.6%; 4.4%) P** = .9740	−0.5% (−6.5%; 5.5%) P** = .9999	
Male vs. female	−0.3% (−9.1%; 8.4%) P** = 1.0000	−0.7% (−8.4%; 6.9%) P** = .9998	−2.3% (−10.3%; 5.7%) P** = .9632				
Age class							.5535~
Q1: ≤26	86.9% (79.8%; 93.0%)[a]	89.8% (83.5%; 95.1%)[a]	83.9% (75.8%; 90.6%)[a]	−3.0% (−16.5%; 10.6%) P** = .9999	−5.9% (−19.3%; 7.5%) P** = .9476	3.0% (−8.8%; 14.7%) P** = .9995	
Q2: 27-36	90.0% (84.0%; 95.1%)[a]	93.7% (89.0%; 98.0%)[a]	91.2% (85.1%; 96.1%)[a]	−3.7% (−12.4%; 5.0%) P** = .9585	−2.5% (−13.6%; 8.6%) P** = .9999	−1.2% (−13.0%; 10.5%) P** = 1.0000	
Q3: 37-52	90.6% (84.8%; 95.3%)[a]	92.9% (87.9%; 97.0%)[a]	95.2% (90.8%; 98.4%)[a]	−2.3% (−11.2%; 6.6%) P** = .9994	2.3% (−6.6%; 11.3%) P** = .9993	−4.6% (−13.5%; 4.2%) P** = .8476	
Q4: ≥53	93.2% (89.2%; 96.8%)[a]	94.1% (90.2%; 97.4%)[a]	94.9% (91.1%; 97.9%)[a]	−0.8% (−6.9%; 5.2%) P** = 1.0000	0.8% (−6.3%; 8.0%) P** = 1.0000	−1.7% (−10.2%; 6.9%) P** = 1.0000	
Q2: 27-36 vs. Q1: ≤26	3.1% (−13.9%; 20.1%) P** = 1.0000	3.9% (−10.7%; 18.4%) P** = .9991	7.3% (−10.2%; 24.8%) P** = .9645				
Q2: 27-36 vs. Q3: 37-52	−0.6% (−15.4%; 14.1%) P** = 1.0000	0.8% (−11.6%; 13.2%) P** = 1.0000	−4.0% (−16.5%; 8.4%) P** = 0.9953				
Q2: 27-36 vs. Q4: ≥53	−3.3% (−16.3%; 9.7%) P** = 0.9995	−0.4% (−11.5%; 10.8%) P** = 1.0000	−3.7% (−15.7%; 8.2%) P** = 0.9966				

TABLE 3 (Continued)

	Diagnostic Performance: Sensitivity						
	AI Algorithms			Mean Difference			
	SmartUrgence	Rayvolve	BoneView	SmartUrgencevs. Rayvolve	BoneView vs. Rayvolve	SmartUrgencevs. BoneView	*P* Value*
Q3: 37-52 vs. Q1: ≤26	3.8% (–12.8%; 20.3%) *P*** = .9998	3.1% (–11.6%; 17.8%) *P*** = .9999	11.3% (–4.8%; 27.4%) *P*** = .4516				
Q3: 37-52 vs. Q4: ≥53	–2.6% (–15.1%; 9.8%) *P*** = .9999	–1.2% (–12.5%; 10.1%) *P*** = 1.0000	0.3% (–9.4%; 10.0%) *P*** = 1.0000				
Q4: ≥53 vs. Q1: ≤26	6.4% (–8.6%; 21.4%) *P*** = 0.9593	4.3% (–9.4%; 17.9%) *P*** = 0.9964	11.0% (–4.7%; 26.7%) *P*** = .4574				
Body region							.3234~
Foot	85.4% (73.1%; 90.2%)[a]	90.8% (83.0%; 95.0%)[a]	98.1% (94.4%; 98.3%)[a]	–5.4% (–51.2%; 40.4%) *P*** = .9269	7.3% (–26.7%; 41.3%) *P*** = .6252	–12.7% (–56.4%; 31.0%) *P*** = .4308	
Knee	68.2% (32.4%; 88.6%)[a]	68.2% (31.8%; 81.4%)[a]	76.7% (36.5%; 89.6%)[a]	0.0% (–100%; 100%) *P*** = 1.0000	8.5% (–63.0%; 80.0%) *P*** = .9251	–8.5% (–100%; 100%) *P*** = .9968	
Ankle	89.9% (76.0%; 95.5%)[a]	92.1% (82.2%; 96.5%)[a]	89.9% (79.4%; 95.4%)[a]	–2.2% (–54.3%; 50.0%) *P*** = .9999	–2.2% (–25.3%; 21.0%) *P*** = .9739	0.0% (–46.7%; 46.7%) *P*** = 1.0000	
Wrist/hand	93.6% (88.0%; 96.3%)[a]	97.8% (94.5%; 99.0%)[a]	91.5% (84.9%; 95.6%)[a]	–4.2% (–24.0%; 15.5%) *P*** = .6269	–6.3% (–34.2%; 21.6%) *P*** = .5888	2.1% (–22.6%; 26.8%) *P*** = .9854	
Knee vs. foot	–17.2% (–100%; 100%) *P*** = .9409	–22.6% (–100%; 100%) *P*** = .8025	–21.4% (–100%; 100%) *P*** = .7773				
Knee vs. ankle	–21.8% (–100%; 100%) *P*** = .8916	–23.9% (–100%; 100%) *P*** = .8142	–13.3% (–100%; 100%) *P*** = .9804				
Knee vs. wrist/hand	–25.4% (–100%; 100%) *P*** = .7580	–29.6% (–100%; 100%) *P*** = .6144	–14.8% (–100%; 100%) *P*** = .9312				
Ankle vs. foot	4.5% (–65.2%; 74.3%) *P*** = .9970	1.3% (–58.9%; 61.5%) *P*** = 1.0000	–8.1% (–62.6%; 46.3%) *P*** = .8360				

TABLE 3 (Continued)

	Diagnostic Performance: Sensitivity						
	AI Algorithms			Mean Difference			
	SmartUrgence	Rayvolve	BoneView	SmartUrgencevs. Rayvolve	BoneView vs. Rayvolve	SmartUrgencevs. BoneView	*P* Value*
Ankle vs. wrist/hand	−3.6% (−60.4%; 53.1%) *P*** = .9974	−5.7% (−54.2%; 42.9%) *P*** = .9309	−1.5% (−59.5%; 56.5%) *P*** = 1.0000				
Wrist/hand vs. foot	8.2% (−44.3%; 60.6%) *P*** = .8151	7.0% (−33.6%; 47.6%) *P*** = .7620	−6.6% (−39.2%; 25.9%) *P*** = .6575				

[a]95% confidence intervals using bootstrap bias-corrected and accelerated method.

AI, artificial intelligence.

* Type III *P* values.

** *P* value adjustment for multiple comparisons using the simulation-based method.

~ *P* values for interaction between AI algorithms and subgroups.

Performance of Each AI Algorithm by Sex, Age, and Body Region and Two-by-Two Comparisons of the Performance of the Algorithms by Subgroups

Results are reported in Tables 2–5. For SmartUrgence, there was no significant difference between subgroups in accuracy, specificity, or sensitivity. For Rayvolve, accuracy and specificity were significantly higher with age 26-37 than ≥53 years (*P* = .0029 and *P* = .0019). Specificity was higher for the subgroup knee than foot (*P* = .0149). For BoneView, accuracy was significantly higher for the subgroups knee than foot (*P* = .0006) and knee than wrist/hand (*P* = .0228). Specificity was significantly higher for the subgroups knee than foot (*P* = .0003) and ankle than foot (*P* = .0195).

Two-by-two comparisons of accuracy and specificity between the three algorithms according to the subgroups of sex, age, and body regions demonstrated significant differences between SmartUrgence and BoneView versus Rayvolve, with no significant difference between SmartUrgence and BoneView. The three algorithms did not differ in sensitivity according to subgroups.

DISCUSSION

We assessed the performance of three commercially available AI algorithms designed for automatic detection of acute fracture with 1500 radiography sets obtained from 1210 consecutive patients admitted to our emergency department for acute skeletal trauma. There were a total of 356 fractured regions, representing 23.7% of examined regions. We found that AI detection of acute peripheral fractures in daily radiological practice in an emergency department is effective and related to patient age, body region, and AI algorithm but not sex.

In real-world conditions, with 1500 unselected consecutive radiography sets, we observed high accuracy (90.1% and 88.8%) and specificity (92.5% and 90.5%) for SmartUrgence and BoneView, respectively. The performances were similarly robust and did not significantly vary by sex or age. These results are comparable to those obtained in previous studies evaluating the performance of commercial algorithms with a known fracture prevalence set at 50% and selection of fractures and high-quality radiographs (9–11). In the study by Duron et al (9), six radiologists and six emergency physicians were asked to detect and localize fractures with and without AI aid (BoneView) from 600 patients, 300 with fractures and 300 without fracture, in six body regions (50 cases with fracture and 50 with no fracture per body region). The stand-alone area under the receiver operating characteristic curve (AUC) of the AI algorithm was 0.91. The AI aid provided a gain of specificity (4.1% increase) and sensitivity (8.7% increase). The mean reading time was reduced by 15.0%. In the study by Guermazi et al (10), of 480 patients with at least 60 examinations per body region (*n* = 8), a fracture prevalence set at 50%, and six types of readers, the AUC was 0.97 for the stand-alone performance of the AI algorithm (BoneView) for fracture detection. AI-assisted radiographic interpretation conferred a 10.4% improvement in fracture detection sensitivity without

TABLE 4. Diagnostic Performance, Specificity (1500 Radiography Sets; Subgroup Sex; Subgroup Age; Subgroup Body Region)

	Diagnostic Performance: Specificity						
	AI Algorithms			Mean Difference			
	SmartUrgence	Rayvolve	BoneView	SmartUrgencevs. Rayvolve	BoneViewvs. Rayvolve	SmartUrgencevs. BoneView	*P* Value*
Specificity							**< .0001**
All radiography sets	92.5% (91.1%; 94.0%)[a]	70.4% (68.1%; 73.0%)[a]	90.5% (89.1%; 92.3%)[a]	22.1% (18.9%; 25.3%) ***P*** ** **< .0001**	20.1% (16.9%; 23.4%) ***P*** < .0001**	2.0% (−0.2%; 4.2%) *P*** = .0790	
Gender							.8226~
Female	93.5% (91.1%; 95.6%)[a]	71.1% (66.9%; 75.5%)[a]	92.0% (89.7%; 94.5%)[a]	22.4% (16.1%; 28.6%) ***P*** < .0001**	20.9% (14.7%; 27.1%) ***P*** < .0001**	1.5% (−2.8%; 5.8%) *P*** = .9240	
Male	92.1% (90.4%; 93.8%)[a]	70.3% (67.3%; 73.6%)[a]	89.7% (87.9%; 91.8%)[a]	21.7% (16.8%; 26.7%) ***P*** < .0001**	19.4% (14.3%; 24.4%) ***P*** < .0001**	2.4% (−1.0%; 5.7%) *P*** = .3188	
Male vs. female	−1.4% (−6.0%; 3.2%) *P*** = .9527	−0.8% (−8.8%; 7.3%) *P*** = .9998	−2.3% (−7.3%; 2.7%) *P*** = .7711				
Age class							**0.0055~**
Q1: ≤26	92.3% (89.7%; 95.3%)[a]	71.6% (67.3%; 76.6%)[a]	93.0% (90.4%; 95.9%)[a]	20.7% (12.2%; 29.2%) ***P*** < .0001**	21.4% (13.2%; 29.6%) ***P*** < .0001**	−0.7% (−5.2%; 3.8%) *P*** = 1.0000	
Q2: 27-36	93.7% (91.3%; 96.5%)[a]	78.9% (74.9%; 83.5%)[a]	92.7% (90.3% 95.2%)[a]	14.8% (7.0%; 22.6%) ***P*** < .0001**	13.8% (5.0%; 22.5%) ***P*** < .0001**	1.0% (−4.8%; 6.8%) *P*** = 1.0000	
Q3: 37-52	93.3% (90.7%; 95.9%)[a]	68.2% (63.7%; 73.3%)[a]	88.0% (85.1%; 91.6%)[a]	25.1% (15.7%; 34.4%) ***P*** < .0001**	19.7% (11.1%; 28.4%) ***P*** < .0001**	5.4% (−1.3%; 12.0%) *P*** = .2548	
Q4: ≥53	90.8% (87.6%; 94.2%)[a]	62.9% (57.5%; 68.2%)[a]	88.4% (84.8%; 92.1%)[a]	27.9% (18.1%; 37.7%) ***P*** < .0001**	25.5% (15.3%; 35.7%) ***P*** < .0001**	2.4% (−4.6%; 9.4%) *P*** = .9927	
Q2: 27-36 vs. Q1: ≤26	1.3% (−5.8%; 8.5%)	7.3% (−4.6%; 19.2%)	-0.4% (−7.4%; 6.6%)				

TABLE 4 (Continued)

	Diagnostic Performance: Specificity						
	AI Algorithms			Mean Difference			
	SmartUrgence	Rayvolve	BoneView	SmartUrgencevs. Rayvolve	BoneViewvs. Rayvolve	SmartUrgencevs. BoneView	*P* Value[*]
	*P*** = 1.0000	*P*** = .6817	*P*** = 1.0000				
Q2: 27-36 vs. Q3: 37-52	0.4% (–6.4%; 7.1%)	10.7% (–1.3%; 22.7%)	4.7% (–3.1%; 12.6%)				
	*P*** = 1.0000	*P*** = .1346	*P*** = .7016				
Q2: 27-36 vs. Q4: ≥53	2.9% (–4.7%; 10.5%)	16.1% (3.6%; 28.6%)	4.3% (–3.8%; 12.4%)				
	*P*** = .9828	***P*** = .0019**	*P*** = .8364				
Q3: 37-52 vs. Q1: ≤26	1.0% (–6.1%; 8.1%)	–3.4% (–16.2%; 9.3%)	–5.1% (–13.2%; 3.0%)				
	*P*** = 1.0000	*P*** = .9992	*P*** = .6327				
Q3: 37-52 vs. Q4: ≥53	2.5% (–5.0%; 10.1%)	5.4% (–7.9%; 18.7%)	–0.4% (–9.4%; 8.6%)				
	*P*** = .9937	*P*** = .9739	*P*** = 1.0000				
Q4: ≥53 vs. Q1: ≤26	–1.6% (–9.5%; 6.3%)	–8.8% (–22.0%; 4.4%)	–4.7% (–13.0%; 3.6%)				
	*P*** = 1.0000	*P*** = .5486	*P*** = .7792				
Body region							0.0868~
Foot	90.4% (87.0%; 95.8%)[a]	62.5% (56.9%; 71.7%)[a]	79.5% (74.4%; 88.0%)[a]	27.9% (13.7%; 42.1%)	17.0% (3.3%; 30.8%)	10.9% (–1.4%; 23.2%)	
				P < .0001**	***P*** = .0040**	*P*** = .1316	
Knee	96.7% (94.6%; 99.2%)[a]	81.5% (76.9%; 87.6%)[a]	97.3% (95.0%; 99.4%)[a]	15.2% (6.0%; 24.4%)	15.8% (6.1%; 25.4%)	–0.5% (–5.9%; 4.8%)	
				P < .0001**	***P*** < .0001**	*P*** = 1.0000	
Ankle	92.0% (88.5%; 96.0%)[a]	72.1% (66.9%; 79.1%)[a]	93.1% (89.9%; 97.1%)[a]	19.9% (9.3%; 30.5%)	21.0% (10.5%; 31.4%)	–1.1% (–8.8%; 6.6%)	
				P < .0001**	***P*** < .0001**	*P*** = 1.0000	
Wrist/hand	91.7% (89.0%; 95.3%)[a]	74.6% (70.1%; 80.9%)[a]	92.1% (89.2%; 95.8%)[a]	17.1% (6.9%; 27.2%)	17.5% (8.1%; 27.0%)	–0.5% (–8.0%; 7.1%)	
				P < .0001**	***P*** < .0001**	*P*** = 1.0000	

TABLE 4 (Continued)

	Diagnostic Performance: Specificity						
	AI Algorithms			Mean Difference			
	SmartUrgence	Rayvolve	BoneView	SmartUrgencevs. Rayvolve	BoneViewvs. Rayvolve	SmartUrgencevs. BoneView	*P* Value*
Knee vs. foot	6.3% (–3.0%; 15.7%) *P*** = .4982	19.1% (2.1%; 36.0%) ***P*** = .0149**	17.8% (5.7%; 29.8%) ***P*** = .0003**				
Knee vs. ankle	4.7% (–3.0%; 12.5%) *P*** = .6541	9.4% (–5.6%; 24.4%) *P*** = .6183	4.2% (–3.3%; 11.7%) *P*** = .7683				
Knee vs. wrist/hand	5.1% (–2.4%; 12.5%) *P*** = .4913	6.9% (–7.1%; 20.9%) *P*** = .8787	5.2% (–2.2%; 12.5%) *P*** = .4378				
Ankle vs. foot	1.6% (–8.0%; 11.3%) *P*** = 1.0000	9.7% (–6.8%; 26.2%) *P*** = .7117	13.6% (1.2%; 26.0%) ***P*** = .0195**				
Ankle vs. wrist/hand	0.3% (–8.5%; 9.2%) *P*** = 1.0000	–2.5% (–17.0%; 12.1%) *P*** = 1.0000	1.0% (–7.5%; 9.5%) *P*** = 1.0000				
Wrist/hand vs. foot	1.3% (–8.9%; 11.5%) *P*** = 1.0000	12.1% (–5.0%; 29.2%) *P*** = .4275	12.6% (–0.2%; 25.5%) *P*** = .0574				

[a] 95% confidence intervals using bootstrap bias-corrected and accelerated method.

AI, artificial intelligence.

[*] Type III *P* values.

[**] *P* value adjustment for multiple comparisons using the simulation-based method.

[~] *P* values for interaction between AI algorithms and subgroups.

TABLE 5. Summary of Diagnostic Performance (Significant Results From Tables 2–4)

	AI Algorithms			Mean Difference		
	SmartUrgence	Rayvolve	BoneView	SmartUrgence vs. Rayvolve	BoneView vs. Rayvolve	SmartUrgence- vs. BoneView
Accuracy						
All	90.1%	71.0%	88.8%	19.2% **P^{**} < .0001**	17.9% **P^{**} < .0001**	1.3% P^{**} = .32
Age class						
Q2: 27-36 vs. Q4:≥ 53	2.7% P^{**} = .9892	13.5% ***P^{**} = .0029***	2.5% P^{**} = .9953			
Body region						
Foot	87.0%	66.3%	82.1%	20.7% **P^{**} < .0001**	15.8% **P^{**} = .0004**	4.9% P^{**} = .9203
Knee	94.3%	80.6%	96.4%	13.8% **P^{**} = .0002**	15.8% **P^{**} < *.0001***	–2.0% P^{**} = .9939
Knee vs. foot	7.3% P^{**} = .3423	14.3% P^{**} = .0841	14.3% **P^{**} = .0006**			
Knee vs. wrist/hand	5.9% P^{**} = .3769	7.8% P^{**} = .6695	8.0% **P^{**} = .0228**			
Specificity						
All	92.5%	70.4%	90.5%	22.1% **P^{**} < .0001**	20.1% **P^{**} < .0001**	2.0% P^{**} = .0790
Age class						
Q2: 27-36 vs. Q4: ≥ 53	2.9% P^{**} = .9828	16.1% **P^{**} = .0019**	4.3% P^{**} = .8364			
Body region						
Foot	90.4%	62.5%	79.5%	27.9% **P^{**} < .0001**	17.0% **P^{**} = .0040**	10.9% P^{**} = .1316
Knee	96.7%	81.5%	97.3%	15.2% **P^{**} < .0001**	15.8% **P^{**} < .0001**	–0.5% P^{**} = 1.0000
Knee vs. foot	6.3% P^{**} = .4982	19.1% **P^{**} = .0149**	17.8% **P^{**} = .0003**			
Ankle vs. foot	1.6% P^{**} = 1.0000	9.7% P^{**} = .7117	13.6% **P^{**} = .0195**			

TABLE 5 (Continued)

	AI Algorithms			Mean Difference		
	SmartUrgence	Rayvolve	BoneView	SmartUrgence vs. Rayvolve	BoneView vs. Rayvolve	SmartUrgence- vs. BoneView
Sensitivity						
All	90.2%	92.6%	91.3%	−2.5% *P*** = .2343	−1.3% *P*** = .6978	−1.1% *P*** = .7554

AI, artificial intelligence.

** *P* value adjustment for multiple comparisons using the simulation-based method.

reducing specificity. A prospective multireader diagnostic accuracy study challenged the performance of the AI algorithm SmartUrgence with 300 radiography sets (skeletal, including skull and spine, chest, and abdomen) from adults and children, approximately half containing one abnormality. The study aimed to determine whether the AI algorithm could pass the radiographic reporting component of the Fellowship of the Royal College of Radiologists examination (11). The AI algorithm achieved an overall average accuracy of 79.5%, and the 26 radiologists an accuracy of 84.1%. In a retrospective, monocentric and observational study including 1772 patients who underwent emergency radiography (skeletal and chest), the overall AUC was 0.95 for SmartUrgence, with no difference across age or body-part subgroups (14). In our study, accuracy (71.0%) and specificity (70.4%) were significantly lower for Rayvolve than SmartUrgence and BoneView. These results differ from those obtained with Rayvolve and 2634 radiography sets from real-life cohort of 2549 children (mean age, 8.5 years; age range: 0-17 years) presenting routinely to the emergency room. Rayvolve yielded 90.4% accuracy and 88.8% specificity (12). The discrepancies could be related to the mean age of the study populations.

For the three algorithms, accuracy and specificity varied by body region, significantly for BoneView and Rayvolve. The highest accuracies and specificities were at the knee and the lowest at the foot. This observation can be explained by the high number of fractures at the foot (30.1% of the radiography sets), the complexity of the anatomy of the foot, normal variants (accessory ossicles), variety of possible injuries and fractures, and the small size of some avulsion fragments or bone impactions. This was not mentioned in previous studies (8–12,14). Our results indicate that the foot requires special attention for the radiologist with or without AI aid. They justify a particular training of algorithms with a large number of normal and fractured cases for that region, ideally with the help of CT images.

In real-world conditions, with the 1500 radiography sets, we observed excellent sensitivity for the three AI algorithms (from 90.2% for SmartUrgence to 92.6% for Rayvolve). These sensitivities are similar to those reported in series with selected radiographs and body regions, and fracture prevalence set at 50% (9–11) or in more real-life conditions (12,14). This is a highly interesting point in that current commercial algorithms can reliably predict a negative radiograph and could serve as a triage tool in the emergency workflow. We had defined the sensitivity as the ability of the AI algorithm to detect at least one existing fracture among the radiography sets with at least one fracture even with some additional incorrect fracture diagnoses. However, we also used a five-category classification to rate the AI algorithm reports. The classification intended to encompass AI false-negative as well as false-positive identifications. The idea was that for the patient a missed fracture can have deleterious consequences. However, a false-positive identification of a fracture can also have deleterious consequences if immobilization is performed. Indeed, an AI false-positive identification is not always corrected by the radiologist. Furthermore, although in many cases AI false-positive identifications can be corrected easily by the practitioner, they provide a loss of confidence in the algorithm, which can lead to subsequent false-negative interpretations by the radiologist (1). Two algorithms provided false-positive identifications in 7.0%-8.7% of radiography sets and the third one in one quarter of cases (26.5%), a substantially higher percentage. Discriminating between a true- and false-positive identification is a major challenge in AI software designed for fracture detection or other tasks of detection. Indeed, current AI algorithms reduce perceptual errors by helping the radiologist detect abnormalities (high sensitivity), for example, fractures in a corner of the radiography or multiple fractures. However, AI assistance should not add cognitive errors that are detections not interpreted correctly (6,15,16). AI algorithms should evolve toward high specificity to be reliable partners of the radiologist.

Although center-dependent, our study provides some epidemiological data on fractures. First, approximately one quarter of patients (27.7%) who had radiographs for a peripheral skeletal trauma had one or more fractured regions. This proportion is interesting information. Indeed, these data on the proportion of fractures among patients undergoing radiography are usually not available from adult fracture epidemiology reports (17–20). Such a proportion highlights one more time that fracture detection is a high-demand reading that could benefit from AI algorithms. Second, more than 60% of patients who underwent radiography were males, corresponding to a ratio of males to females of 1.6, but the ratio of males to females with fracture was only 1.2. So, there were more fractures in women than in men undergoing radiography. Another example concerns the body regions fractured. In our series from a general emergency department, two thirds of fractures were in the upper extremity and girdle. Therefore, series selecting cases, for example, with equivalent numbers of cases for the regions shoulder, arm, hand, pelvis, leg, and foot, inevitably introduced bias.

Our study has several limitations. First, the data set was from one single center, which could limit the generalizability of the results. In fact, the population of Lariboisière hospital, Paris (Assistance Publique-Hôpitaux de Paris-Université Paris Cité), particularly in the north of Lariboisière hospital, Paris (Assistance Publique-Hôpitaux de Paris-Université Paris Cité), where our hospital is located, is cosmopolitan, consisting of people of different origins. The radiography set was quite large and represents the whole range of possible patients aged ≥15 years with a peripheral skeletal trauma, which limits spectrum bias. Some variability was also present because 30 radiographers were involved day and night in radiographic acquisitions. Second, the ground truth can be seen as weak compared with systematic CT evaluation, but radiation exposure and emergency workflow are not compatible with such consideration. Third, we cannot evaluate whether the AI algorithms aided the emergency physician in the context of patient care because the study was not designed for such a complex question. Finally, the study provides the evaluation of three AI algorithms at a given time, but these algorithms are becoming more efficient every day, evolving toward more accuracy in fracture detection, and a wider range of detected items such as dislocation and joint fluid, and a more exhaustive skeletal evaluation including ribs and spine.

Among the strengths of our study is the consecutive inclusion of radiographs, without selecting patients, body regions, fractures, or high-quality radiographs. Second, the algorithms were not previously trained on radiographs from our department, which reflects the real-life use of commercial AI algorithms. This point is to be highlighted because diagnostic performance is known to decrease in external data sets, sometimes greatly (21), compared with internal validations (21,22). Third, the quite large study population allowed us to perform statistical analyses by subgroups and underline non-uniform results between subgroups. Finally, the three algorithms were evaluated for 13 body regions in the same radiography collection. This is a unique experience to the best of our knowledge. The similarity of the performance achieved with two of the algorithms adds confidence to the findings and provides some references for daily radiological practice.

To conclude, the role of deep learning and its application in radiology practice is evolving. Important challenges remain, such as the diffusion and validation of these algorithms in daily radiological practice. In our study, we evaluated three AI algorithms for fracture detection of 13 body regions in the daily radiological practice of one emergency department. The performance of the algorithms was good to high. The variability in performance was related to the algorithm, the body region examined, and age but not sex. The directions that could be followed to further augment AI for fracture detection in clinical workflow would be to focus the training on anatomical locations frequently experiencing trauma and for which the algorithms show weaknesses, such as the foot region. Also, all commercial algorithms should become able to propose an exhaustive evaluation with the analysis of the ribs, the spine, and the craniofacial skeleton, especially for use in countries with limited access to CT scanners.

FUNDING

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

DECLARATION OF COMPETING INTEREST

Valérie Bousson is a paid consultant for Milvue starting July 1, 2022. Grégoire Attané was paid by Gleamer between March and August 2019 to label radiographs. The remaining authors declare none. The three companies, Milvue, AZmed, and Gleamer, provided the AI algorithms to our radiology department for free. None of the three companies had access to the study methodology or data during the course of the study.

ACKNOWLEDGMENTS

We are grateful to the radiographers who manually transferred all radiographs for evaluation by each AI algorithm and checked the availability of AI reports on the PACS, especially Lecolazet Didiane, Rota Agathe, Caparos Alain, and Barat Stephane; the six residents involved in the preparatory step (Assouline Victoria, Barberis Eric, Beunon Paul, Chanclud Justine, Chung Cecile, and Kedra Alice); the Information Technology and the Medical Physic Departments of Lariboisiere; and the companies AZmed, Gleamer, and Milvue that provided free AI algorithms to our radiology department.

APPENDIX A

See Tables A1 and A2.

TABLE A1. Descriptive Statistics: Classification of AI Algorithms Results (vs. Final Diagnosis by Senior Radiologists), Whole Population

	Radiography sets N = 1500
Results of AI algorithm (vs. senior radiologists as gold standard, presence/absence of single or multiple fractures), SmartUrgence, no. (%)	
A. True-positive and/or true-negative identifications	1356 (90.4%)
B. True-positive plus false-positive identifications	22 (1.5%)
C. False-positive identification	82 (5.5%)
D. True-positive plus false-negative identifications	7 (0.5%)
E. False-negative identification	33 (2.2%)
All	1500 (100.0%)
Results of AI algorithm (vs. senior radiologists as gold standard, presence/absence of single or multiple fractures), Rayvolve, no. (%)	
A. True-positive and/or true-negative identifications	1070 (71.3%)
B. True-positive plus false-positive identifications	66 (4.4%)
C. False-positive identification	331 (22.1%)
D. True-positive plus false-negative identifications	8 (0.5%)
E. False-negative identification	25 (1.7%)
All	1500 (100.0%)
Results of AI algorithm (vs. senior radiologists as gold standard, presence/absence of single or multiple fractures), BoneView, no. (%)	
A. True-positive and/or true-negative identifications	1336 (89.1%)
B. True-positive plus false-positive identifications	26 (1.7%)
C. False-positive identification	105 (7.0%)
D. True-positive plus false-negative identifications	5 (0.3%)
E. False-negative identification	28 (1.9%)
All	1500 (100.0%)

Data are number (%) of radiography sets.
AI, artificial intelligence.

TABLE A2. Descriptive Statistics: Classification of AI Algorithms Results (vs. Final Diagnosis by Senior Radiologists) According to (a) Sex, (b) Age Class, and (c) Body Region (Four Most Commonly Examined)

	Female	Male	All	*P* Value
(a)				
Final diagnosis of senior radiologist, no. (%)				
Fracture	162 (27.7%)	194 (21.2%)	356 (23.7%)	
No fracture	422 (72.3%)	722 (78.8%)	1144 (76.3%)	
All	584 (100.0%)	916 (100.0%)	1500 (100.0%)	
Results of AI algorithm SmartUrgence, no. (%)				.34
A. True-positive and/or true-negative identifications	532 (91.1%)	824 (90.0%)	1356 (90.4%)	
B. True-positive plus false-positive identifications	10 (1.7%)	12 (1.3%)	22 (1.5%)	
C. False-positive identification	26 (4.5%)	56 (6.1%)	82 (5.5%)	
D. True-positive plus false-negative identifications	1 (0.2%)	6 (0.7%)	7 (0.5%)	
E. False-negative identification	15 (2.6%)	18 (2.0%)	33 (2.2%)	
All	584 (100.0%)	916 (100.0%)	1500 (100.0%)	
Results of AI algorithm Rayvolve, no. (%)				.62
A. True-positive and/or true-negative identifications	424 (72.6%)	646 (70.5%)	1070 (71.3%)	
B. True-positive plus false-positive identifications	28 (4.8%)	38 (4.1%)	66 (4.4%)	
C. False-positive identification	119 (20.4%)	212 (23.1%)	331 (22.1%)	
D. True-positive plus false-negative identifications	2 (0.3%)	6 (0.7%)	8 (0.5%)	
E. False-negative identification	11 (1.9%)	14 (1.5%)	25 (1.7%)	
All	584 (100.0%)	916 (100.0%)	1500 (100.0%)	
Results of AI algorithm BoneView, no. (%)				.48
A. True-positive and/or true-negative identifications	529 (90.6%)	807 (88.1%)	1336 (89.1%)	
B. True-positive plus false-positive identifications	10 (1.7%)	16 (1.7%)	26 (1.7%)	
C. False-positive identification	32 (5.5%)	73 (8.0%)	105 (7.0%)	
D. True-positive plus false-negative identifications	2 (0.3%)	3 (0.3%)	5 (0.3%)	
E. False-negative identification	11 (1.9%)	17 (1.9%)	28 (1.9%)	
All	584 (100.0%)	916 (100.0%)	1500 (100.0%)	

	Q1: ≤26	Q2: 27-36	Q3: 37-52	Q4: ≥53	All	*P* Value
(b)						
Imaging result of senior radiologist, no. (%)						
Fracture	67 (18.8%)	82 (21.6%)	88 (22.7%)	119 (31.5%)	356 (23.7%)	
No fracture	289 (81.2%)	297 (78.4%)	299 (77.3%)	259 (68.5%)	1144 (76.3%)	
All	356 (100.0%)	379 (100.0%)	387 (100.0%)	378 (100.0%)	1500 (100.0%)	
Results of AI algorithm SmartUrgence, no. (%)						.56
A. True-positive and/or true-negative identifications	322 (90.4%)	343 (90.5%)	356 (92.0%)	335 (88.6%)	1356 (90.4%)	
B. True-positive plus false-positive identifications	4 (1.1%)	6 (1.6%)	2 (0.5%)	10 (2.6%)	22 (1.5%)	
C. False-positive identification	21 (5.9%)	19 (5.0%)	20 (5.2%)	22 (5.8%)	82 (5.5%)	
D. True-positive plus false-negative identifications	0 (0.0%)	3 (0.8%)	1 (0.3%)	3 (0.8%)	7 (0.5%)	
E. False-negative identification	9 (2.5%)	8 (2.1%)	8 (2.1%)	8 (2.1%)	33 (2.2%)	

TABLE A2 (Continued)

	Q1: ≤26	Q2: 27-36	Q3: 37-52	Q4: ≥53	All	*P* Value
All	356 (100.0%)	379 (100.0%)	387 (100.0%)	378 (100.0%)	1500 (100.0%)	
Results of AI algorithm Rayvolve, no. (%)						<.01
A. True-positive and/or true-negative identifications	260 (73.0%)	292 (77.0%)	273 (70.5%)	245 (64.8%)	1070 (71.3%)	
B. True-positive plus false-positive identifications	9 (2.5%)	15 (4.0%)	12 (3.1%)	30 (7.9%)	66 (4.4%)	
C. False-positive identification	80 (22.5%)	63 (16.6%)	95 (24.5%)	93 (24.6%)	331 (22.1%)	
D. True-positive plus false-negative identifications	0 (0.0%)	4 (1.1%)	1 (0.3%)	3 (0.8%)	8 (0.5%)	
E. False-negative identification	7 (2.0%)	5 (1.3%)	6 (1.6%)	7 (1.9%)	25 (1.7%)	
All	356 (100.0%)	379 (100.0%)	387 (100.0%)	378 (100.0%)	1500 (100.0%)	
Results of AI algorithm BoneView, no. (%)						.19
A. True-positive and/or true-negative identifications	321 (90.2%)	340 (89.7%)	342 (88.4%)	333 (88.1%)	1336 (89.1%)	
B. True-positive plus false-positive identifications	5 (1.4%)	7 (1.8%)	4 (1.0%)	10 (2.6%)	26 (1.7%)	
C. False-positive identification	19 (5.3%)	22 (5.8%)	36 (9.3%)	28 (7.4%)	105 (7.0%)	
D. True-positive plus false-negative identifications	0 (0.0%)	3 (0.8%)	1 (0.3%)	1 (0.3%)	5 (0.3%)	
E. False-negative identification	11 (3.1%)	7 (1.8%)	4 (1.0%)	6 (1.6%)	28 (1.9%)	
All	356 (100.0%)	379 (100.0%)	387 (100.0%)	378 (100.0%)	1500 (100.0%)	

	All	Hand/Wrist	Knee	Ankle	Foot	*P* Value
(c)						
Final diagnosis of senior radiologist, no. (%)						
Fracture	208 (22.4%)	97 (30.9%)	13 (6.6%)	42 (18.1%)	56 (30.1%)	
No fracture	721 (77.6%)	217 (69.1%)	184 (93.4%)	190 (81.9%)	130 (69.9%)	
All	929 (100.0%)	314 (100.0%)	197 (100.0%)	232 (100.0%)	186 (100.0%)	
Results of AI algorithm SmartUrgence, no. (%)						.1761
A. True-positive and/or true-negative identifications	836 (90.0%)	278 (88.5%)	186 (94.4%)	209 (90.1%)	163 (87.6%)	
B. True-positive plus false-positive identifications	15 (1.6%)	9 (2.9%)	0 (0.0%)	3 (1.3%)	3 (1.6%)	
C. False-positive identification	50 (5.4%)	18 (5.7%)	6 (3.0%)	14 (6.0%)	12 (6.5%)	
D. True-positive plus false-negative identifications	6 (0.6%)	3 (1.0%)	2 (1.0%)	1 (0.4%)	0 (0.0%)	
E. False-negative identification	22 (2.4%)	6 (1.9%)	3 (1.5%)	5 (2.2%)	8 (4.3%)	
All	929 (100.0%)	314 (100.0%)	197 (100.0%)	232 (100.0%)	186 (100.0%)	
Results of AI algorithm Rayvolve, no. (%)						.0002
A. True-positive and/or true-negative identifications	675 (72.7%)	229 (72.9%)	159 (80.7%)	162 (69.8%)	125 (67.2%)	
B. True-positive plus false-positive identifications	44 (4.7%)	25 (8.0%)	0 (0.0%)	11 (4.7%)	8 (4.3%)	
C. False-positive identification	188 (20.2%)	55 (17.5%)	34 (17.3%)	51 (22.0%)	48 (25.8%)	
D. True-positive plus false-negative identifications	8 (0.9%)	3 (1.0%)	1 (0.5%)	4 (1.7%)	0 (0.0%)	
E. False-negative identification	14 (1.5%)	2 (0.6%)	3 (1.5%)	4 (1.7%)	5 (2.7%)	
All	929 (100.0%)	314 (100.0%)	197 (100.0%)	232 (100.0%)	186 (100.0%)	

TABLE A2 (Continued)

	All	Hand/Wrist	Knee	Ankle	Foot	*P* Value
Results of AI algorithm BoneView, no. (%)						**<.0001**
A. True-positive and/or true-negative identifications	835 (89.9%)	278 (88.5%)	190 (96.4%)	213 (91.8%)	154 (82.8%)	
B. True-positive plus false-positive identifications	13 (1.4%)	8 (2.5%)	0 (0.0%)	0 (0.0%)	5 (2.7%)	
C. False-positive identification	60 (6.5%)	17 (5.4%)	5 (2.5%)	12 (5.2%)	26 (14.0%)	
D. True-positive plus false-negative identifications	5 (0.5%)	3 (1.0%)	0 (0.0%)	2 (0.9%)	0 (0.0%)	
E. False-negative identification	16 (1.7%)	8 (2.5%)	2 (1.0%)	5 (2.2%)	1 (0.5%)	
All	929 (100.0%)	314 (100.0%)	197 (100.0%)	232 (100.0%)	186 (100.0%)	

AI, artificial intelligence.

Bold values indicates $P < 0.05$.

REFERENCES

1. Doi K. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. Comput Med Imaging Graph 2007; 31(4–5):198–211. https://doi.org/10.1016/j.compmedimag.2007.02.002
2. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017; 42:60–88. https://doi.org/10.1016/j.media.2017.07.005
3. Miller DD, Brown EW. How cognitive machines can augment medical imaging. AJR Am J Roentgenol 2019; 212:9–14. https://doi.org/10.2214/AJR.18.19914
4. Chea P, Mandell JC. Current applications and future directions of deep learning in musculoskeletal radiology. Skeletal Radiol 2020; 49:183–197. https://doi.org/10.1007/s00256-019-03284-z
5. Kijowski R, Liu F, Caliva F, et al. Deep learning for lesion detection, progression, and prediction of musculoskeletal disease. J Magn Reson Imaging 2020; 52:1607–1619. https://doi.org/10.1002/jmri.27001
6. Bruno MA, Walker EA, Abujudeh HH. Understanding and confronting our mistakes: the epidemiology of error in radiology and strategies for error reduction. Radiographics 2015; 35:1668–1676. https://doi.org/10.1148/rg.2015150023
7. Mock C, Cherian MN. The global burden of musculoskeletal injuries: challenges and solutions. Clin Orthop Relat Res 2008; 466:2306–2316. https://doi.org/10.1007/s11999-008-0416-z
8. Kuo RYL, Harrison C, Curran T-A, et al. Artificial intelligence in fracture detection: a systematic review and meta-analysis. Radiology 2022:211785 https://doi.org/10.1148/radiol.211785
9. 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; 300:120–129. https://doi.org/10.1148/radiol.2021203886
10. Guermazi A, Tannoury C, Kompel AJ, et al. Improving radiographic fracture recognition performance and efficiency using artificial intelligence. Radiology 2022; 302:627–636. https://doi.org/10.1148/radiol.210937
11. Shelmerdine SC, Martin H, Shirodkar K, et al. FRCR-AI Study Collaborators. Can artificial intelligence pass the Fellowship of the Royal College of Radiologists examination? Multi-reader diagnostic accuracy study. BMJ 2022; 379:e072826 https://doi.org/10.1136/bmj-2022-072826
12. Dupuis M, Delbos L, Veil R, Adamsbaum C. External validation of a commercially available deep learning algorithm for fracture detection in children. Diagn Interv Imaging 2022; 103:151–159. https://doi.org/10.1016/j.diii.2021.10.007
13. Santomartino SM, Siegel E, Yi PH. Academic radiology departments should lead artificial intelligence initiatives. Acad Radiol 2023; 30:971–974. https://doi.org/10.1016/j.acra.2022.07.011
14. Parpaleix A, Parsy C, Cordari M, et al. Assessment of a combined musculoskeletal and chest deep learning-based detection solution in an emergency setting. Eur J Radiol Open 2023; 10:100482 https://doi.org/10.1016/j.ejro.2023.100482
15. Renfrew DL, Franken EA, Berbaum KS, et al. Error in radiology: classification and lessons in 182 cases presented at a problem case conference. Radiology 1992; 183:145–150. https://doi.org/10.1148/radiology.183.1.1549661
16. Waite S, Scott J, Gale B, Fuchs T, et al. Interpretive error in radiology. AJR Am J Roentgenol 2017; 208:739–749. https://doi.org/10.2214/AJR.16.16963
17. Court-Brown CM, Caesar B. Epidemiology of adult fractures: a review. Injury 2006; 37:691–697. https://doi.org/10.1016/j.injury.2006.04.130
18. Taylor A, Young A. Epidemiology of orthopaedic trauma admissions over one year in a district general hospital in England. Open Orthop J 2015; 9:191–193. https://doi.org/10.2174/1874325001509010191
19. Rosengren BE, Karlsson M, Petersson I, et al. The 21st-century landscape of adult fractures: cohort study of a complete adult regional population. J Bone Miner Res 2015; 30:535–542. https://doi.org/10.1002/jbmr.2370
20. Singer BR, McLauchlan GJ, Robinson CM, et al. Epidemiology of fractures in 15,000 adults: the influence of age and gender. J Bone Joint Surg Br 1998; 80:243–248. https://doi.org/10.1302/0301-620x.80b2.7762
21. Yu AC, Mohajer B, Eng J. External validation of deep learning algorithms for radiologic diagnosis: a systematic review. Radiol Artif Intell 2022; 4:e210064 https://doi.org/10.1148/ryai.210064
22. Mongan J, Moy L, Kahn CE. Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers. Radiol Artif Intell 2020; 2:e200029 https://doi.org/10.1148/ryai.2020200029