Radiology

Artificial Intelligence in Fracture Detection: A Systematic Review and Meta-Analysis

Rachel Y. L. Kuo, MB BChir, MA, MRCS • Conrad Harrison, BSc, MBBS, MRCS • Terry-Ann Curran, MB BCh BAO, MD • Benjamin Jones, BMBCh, BA • Alexander Freethy, BSc, MBBS, MSc, MRCS • David Cussons, BSc, MBBS • Max Stewart, MB BChir, BA • Gary S. Collins, BSc, PhD • Dominic Furniss, DM, MA, MBBCh, FRCS (Plast)

From the Nuffield Department of Orthopedics, Rheumatology and Musculoskeletal Sciences, Botnar Research Centre, Old Road Headington, Oxford OX3 7LD, UK (R.Y.L.K., C.H., M.S., G.S.C., D.F.); Department of Plastic Surgery, John Radcliffe Hospital, Oxford, UK (T.A.C., A.F.); Department of Vascular Surgery, Royal Berkshire Hospital, Reading, UK (B.J.); Department of Plastic Surgery, Stoke Mandeville Hospital, Aylesbury, Buckinghamshire UK (D.C.); and UK EQUATOR Center, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford Centre for Statistics in Medicine, Oxford UK (G.S.C.). Received July 14, 2021; revision requested August 16; revision received January 15, 2022; accepted January 21. **Address correspondence to** R.Y.K. (e-mail: *rachel.kuo@ndorms.ox.ac.uk*).

R.K. supported by a National Institute for Health Research (NIHR) Academic Clinical Fellowship. C.H. supported by a NIHR Doctoral Research Fellowship (NIHR300684). D.F. supported by the Oxford NIHR Biomedical Research Centre. This publication presents independent research funded by the NIHR.

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

See also the editorial by Cohen and McInnes in this issue.

Radiology 2022; 304:50–62 • https://doi.org/10.1148/radiol.211785 • Content codes: IN MK

Background: Patients with fractures are a common emergency presentation and may be misdiagnosed at radiologic imaging. An increasing number of studies apply artificial intelligence (AI) techniques to fracture detection as an adjunct to clinician diagnosis.

Purpose: To perform a systematic review and meta-analysis comparing the diagnostic performance in fracture detection between AI and clinicians in peer-reviewed publications and the gray literature (ie, articles published on preprint repositories).

Materials and Methods: A search of multiple electronic databases between January 2018 and July 2020 (updated June 2021) was performed that included any primary research studies that developed and/or validated AI for the purposes of fracture detection at any imaging modality and excluded studies that evaluated image segmentation algorithms. Meta-analysis with a hierarchical model to calculate pooled sensitivity and specificity was used. Risk of bias was assessed by using a modified Prediction Model Study Risk of Bias Assessment Tool, or PROBAST, checklist.

Results: Included for analysis were 42 studies, with 115 contingency tables extracted from 32 studies (55 061 images). Thirty-seven studies identified fractures on radiographs and five studies identified fractures on CT images. For internal validation test sets, the pooled sensitivity was 92% (95% CI: 88, 93) for AI and 91% (95% CI: 85, 95) for clinicians, and the pooled specificity was 91% (95% CI: 88, 93) for AI and 92% (95% CI: 89, 92) for clinicians. For external validation test sets, the pooled sensitivity was 91% (95% CI: 84, 95) for AI and 94% (95% CI: 90, 96) for clinicians, and the pooled specificity was 91% (95% CI: 81, 95) for AI and 94% (95% CI: 91, 95) for clinicians. There were no statistically significant differences between clinician and AI performance. There were 22 of 42 (52%) studies that were judged to have high risk of bias. Meta-regression identified multiple sources of heterogeneity in the data, including risk of bias and fracture type.

Conclusion: Artificial intelligence (AI) and clinicians had comparable reported diagnostic performance in fracture detection, suggesting that AI technology holds promise as a diagnostic adjunct in future clinical practice.

Clinical trial registration no. CRD42020186641

Online supplemental material is available for this article.

Fractures have an incidence of between 733 and 4017 per 100 000 patient-years (1–3). In the financial year April 2019 to April 2020, 1.2 million patients presented to an emergency department in the United Kingdom with an acute fracture or dislocation, an increase of 23% from the year before (4). Missed or delayed diagnosis of fractures on radiographs is a common diagnostic error, ranging from 3% to 10% (5–7). There is an inverse relationship between clinician experience and rate of fracture misdiagnosis, but timely access to expert opinion is not widely available (6). Growth in imaging volumes continues to outpace radiologist recruitment: A Canadian study (8) from 2019 found an increase in radiologist workloads of 26% over 12 years, whereas a study from the American College of Radiologists found a 30% increase in job openings from 2017 to 2018 (9). Strategies (6,10) to reduce rates of fracture misdiagnosis and to streamline patient pathways are crucial to maintain high standards of patient care.

Artificial intelligence (AI) is a branch of computer science in which algorithms perform tasks traditionally assigned to humans. *Machine learning* is a term that refers to a group of techniques in the field of AI that allow algorithms to learn from data, iteratively improving their own performance without the need for explicit programming. *Deep learning* is a term often used interchangeably with machine learning but refers to algorithms that use multiple processing layers to extract high level information from any input. Health care, and

Abbreviations

AI = artificial intelligence, TRIPOD = Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis

Summary

Artificial intelligence is noninferior to clinicians in terms of diagnostic performance in fracture detection, showing promise as a useful diagnostic tool.

Key Results

- In a systematic review and meta-analysis of 42 studies (37 studies with radiography and five studies with CT), the pooled diagnostic performance from the use of artificial intelligence (AI) to detect fractures had a sensitivity of 92% and 91% and specificity of 91% and 91%, on internal and external validation, respectively.
- Clinician performance had comparable performance to AI in fracture detection (sensitivity 91%, 92%; specificity 94%, 94%).
- Only 13 studies externally validated results, and only one study evaluated AI performance in a prospective clinical trial.

in particular, radiologic image classification, has been identified as a key sector in which AI could streamline pathways, acting as a triage or screening service, as a decision aid, or as second-reader support for radiologists (10).

Recent narrative reviews have reported high accuracy for deep learning in fracture detection and classification. Smets et al (11) summarized 32 studies, finding a wide range of accuracy (78%–99%) similar to Langerhuizen et al (12), who found a range of 77%–90% accuracy across 10 studies. Recent studies reported higher accuracy estimates (93%–99%) (12–14). Yang et al (14) performed a meta-analysis of nine studies with a pooled sensitivity and specificity of 87% and 91%, respectively.

Our study is a systematic review and meta-analysis of 42 studies, comparing the diagnostic performance in fracture detection between AI and clinicians in peer-reviewed publications and in the gray literature (ie, articles published on preprint repositories) on radiographs or CT images. We described study methods, adherence to reporting guidelines, and we performed a detailed assessment of risk of bias and study applicability.

Materials and Methods

Protocol and Registration

This systematic review was prospectively registered with PROSPERO (CRD42020186641). Our study was prepared by using guidelines from the Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies (15,16). All stages of the review (title and abstract screening, full-text screening, data extraction, assessment of adherence to reporting guidelines, bias, and applicability) were performed in

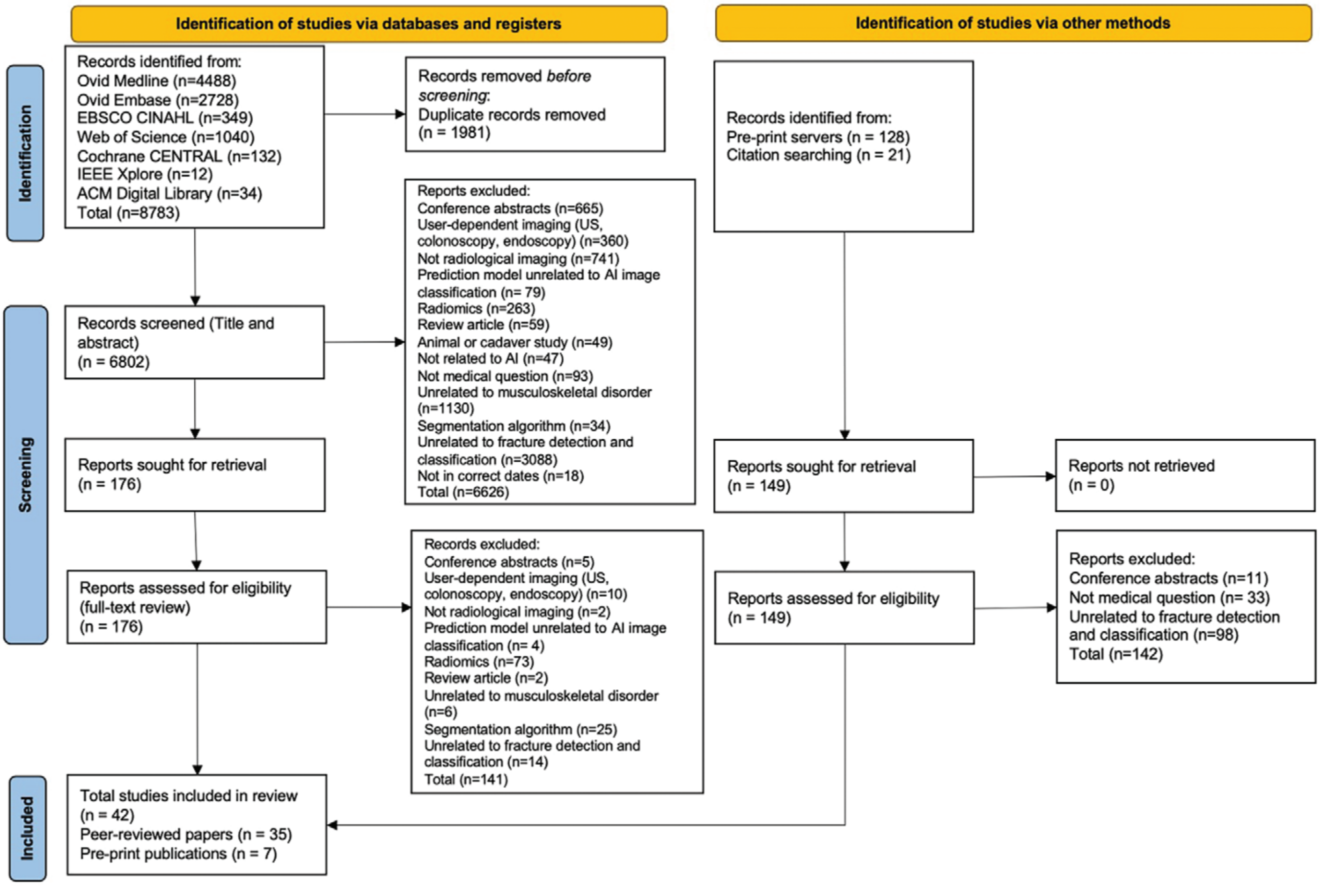
Identification of studies via databases and registers
Identification of studies via other methods
Identification
Screening
Included
Records identified from:
Ovid Medline (n=4488)
Ovid Embase (n=2728)
EBSCO CINAHL (n=349)
Web of Science (n=1040)
Cochrane CENTRAL (n=132)
IEEE Xplore (n=12)
ACM Digital Library (n=34)
Total (n=8783)
Records removed before screening:
Duplicate records removed
(n = 1981)
Records identified from:
Pre-print servers (n = 128)
Citation searching (n = 21)
Records screened (Title and abstract)
(n = 6802)
Reports excluded:
Conference abstracts (n=665)
User-dependent imaging (US, colonoscopy, endoscopy) (n=360)
Not radiological imaging (n=741)
Prediction model unrelated to AI image classification (n= 79)
Radiomics (n=263)
Review article (n=59)
Animal or cadaver study (n=49)
Not related to AI (n=47)
Not medical question (n=93)
Unrelated to musculoskeletal disorder (n=1130)
Segmentation algorithm (n=34)
Unrelated to fracture detection and classification (n=3088)
Not in correct dates (n=18)
Total (n=6626)
Reports sought for retrieval
(n = 176)
Reports sought for retrieval
(n = 149)
Reports not retrieved
(n = 0)
Reports assessed for eligibility (full-text review)
(n = 176)
Records excluded:
Conference abstracts (n=5)
User-dependent imaging (US, colonoscopy, endoscopy) (n=10)
Not radiological imaging (n=2)
Prediction model unrelated to AI image classification (n= 4)
Radiomics (n=73)
Review article (n=2)
Unrelated to musculoskeletal disorder (n=6)
Segmentation algorithm (n=25)
Unrelated to fracture detection and classification (n=14)
Total (n=141)
Reports assessed for eligibility
(n = 149)
Reports excluded:
Conference abstracts (n=11)
Not medical question (n= 33)
Unrelated to fracture detection and classification (n=98)
Total (n=142)
Total studies included in review
(n = 42)
Peer-reviewed papers (n = 35)
Pre-print publications (n = 7)

Figure 1: Preferred Reporting Items for Systematic Reviews and Meta-Analyses flowchart shows studies selected for review. ACM = Association for Computing Machinery, AI = artificial intelligence, CENTRAL = Central Register of Controlled Trials, CINAHL = Cumulative Index to Nursing and Allied Health Literature, IEEE = Institute of Electrical and Electronics Engineers and Institution of Engineering and Technology.

Table 1: Characteristics of Studies, Developing and Internally Validating Algorithms

						No. of Images per Set					
First Author	Year	Imaging Modality	Target Condition	View	Comparison Group	Training	Tuning	Testing	Reference Standard	Model Output	Peer Review Status
With a comparison group											
Adams (46)	2018	Radiography	Proximal femur fracture	AP view	Comparison of two algorithms, nonexpert clinicians	643	...	161	Surgical confirmation	NR	Yes
Chen (35)	2021	Radiography	Vertebral fractures	Frontal view	Expert clinicians	1045	N	261	Expert consensus	Binary classification and saliency map	Yes
Chung* (53)	2018	Radiography	Upper humerus fracture	AP view	Expert clinicians	NR	…	…	Expert consensus	NR	Yes
Gan (51)	2019	Radiography	Distal radius fracture	AP view	Expert clinicians	5202	918	300	Expert consensus	Binary classification	Yes
Jimenez-Sanchez (50)	2020	Radiography	Proximal femur fractures	AP view	Expert clinicians	943	135	269	Expert consensus	NR	Yes
Kim (58)	2018	Radiography	Distal radius or ulna fracture	Lateral view	Expert clinicians	8890	1111	1111	Single expert opinion	Probability of fracture	Yes
Krogue (30)	2020	Radiography	Proximal femur fractures	AP view	Expert clinicians with and without algorithm assistance	1849	739	438	Nonexpert consensus, with reference to other imaging in cases of uncertainty	Probability of fracture and saliency map	Yes
Langerhuizen (55)	2020	Radiography	Scaphoid fracture	Scaphoid series	Expert clinicians	180	20	100	MRI report	Probability of fracture	Yes
Mawatari (29)	2020	Radiography	Proximal femur fractures	AP view	Expert and nonexpert clinicians, with and without algorithm assistance	550	N	50	Expert consensus, using CT/MRI for reference	Probability of fracture	Yes
Murata[†] (28)	2020	Radiography	Vertebral fractures	AP and lateral view	Expert and nonexpert clinicians	NR	…	…	MRI report	NR	Yes
Ozkaya (27)	2020	Radiography	Scaphoid fracture	AP view	Expert and nonexpert clinicians	203	87	100	CT report and single expert opinion	NR	Yes
Pranata (62)	2019	CT	Calcaneal fractures	NR	Comparison of multiple algorithms	1550	N	381	Radiological report	NR	Yes

Table 1 (*continues*)

Table 1 (*continued*): Characteristics of Studies, Developing and Internally Validating Algorithms

First Author	Year	Imaging Modality	Target Condition	View	Comparison Group	No. of Images per Set			Reference Standard	Model Output	Peer Review Status
						Training	Tuning	Testing			
Raisuddin[‡] (24)	2020	Radiography	Distal radius fracture	Concatenated AP and lateral view	Expert and nonexpert clinicians	1946	N	N	Expert consensus, with CT verification in "Test set 2"	Probability of fracture and saliency map	No
Urakawa (43)	2018	Radiography	Intertrochanteric proximal femur fractures	AP view	Expert clinicians	2678	334	334	Single expert opinion	Binary classification	Yes
Yamada (26)	2020	Radiography	Proximal femur fractures	Separate and combined AP/lateral view	Expert clinicians	2632	N	300	Expert consensus and CT/MRI results	NR	Yes
Yu[§] (42)	2020	Radiography	Proximal femur fracture	AP view	Expert clinicians	637	212	212	Radiological report	Binary classification with bounding box	Yes
Without a comparison group											
Beyaz[‖] (40)	2020	Radiography	Proximal femur fracture	AP view	None	NR	…	…	NR	NR	Yes
Derkatch (52)	2019	Radiography		Vertebral fractures, lateral view	None	7646	1274	3822	Expert consensus	Binary classification and saliency map	Yes
Grauhan (57)	2021	Radiography	Proximal humerus fracture	Unspecified views	None	2700	675	269	Single expert opinion, with expert consensus for test set	Probability of fracture and saliency map	Yes
Mehta (47)	2019	Radiography	L1–4 vertebral fractures	AP view	None	246	N	61	Expert consensus	NR	Yes
Mutasa (48)	2020	Radiography	Proximal femur fractures	AP view	None	7250	N	1813	Single expert opinion	Probability of fracture and saliency map	Yes
Raghavendra (61)	2018	CT	T11-L1 vertebral fractures	Sagittal view	None	783	N	336	Single expert opinion	NR	Yes
Rayan (45)	2019	Radiography	Supracondylar or lateral condyle elbow fracture	AP or lateral view	None	20 350	N	3096	Radiological report, single expert opinion in test set	Probability of fracture	Yes
Sato (64)	2020	Radiography	Proximal femur fractures	AP view	None	8484	1000	1000	Expert consensus	Probability of fracture and saliency map	No

Table 1 (*continues*)

Table 1 (*continued*): Characteristics of Studies, Developing and Internally Validating Algorithms

						No. of Images per Set					
First Author	Year	Imaging Modality	Target Condition	View	Comparison Group	Training	Tuning	Testing	Reference Standard	Model Output	Peer Review Status
Starosolski (49)	2019	Radiography	Tibial fracture	AP or lateral view	None	784	98	98	Radiological report	Probability of fracture and saliency map	No
Yahalomi (41)	2018	Radiography	Distal radius fracture	AP view	None	3583	N	893	Single expert opinion	Binary classification with bounding box	No
Yoon (23)	2021	Radiography	Scaphoid fracture	AP and ulnar deviated views	None	8356	1177	2305	Expert consensus	Probability of fracture and saliency map	Yes

Note.—AP = anteroposterior, N = no tuning set, NR = not reported.
* Ten-fold cross-validation (*n* = 189).
† Fivefold cross-validation (*n* = 300).
‡ Test set 1, 207 images; test set 2, 105 images.
§ Twenty-fold cross validation.
‖ Fivefold cross-validation (*n* = 2106).

duplicate by two independent reviewers (R.Y.L.K. and either C.H., T.A.C., B.J., A.F., D.C., or M.S.), and disagreements were resolved by discussion with a third independent reviewer (G.S.C. or D.F.).

Search Strategy and Study Selection

A search was performed to identify studies that developed and/or validated an AI algorithm for the purposes of fracture detection. A search strategy was developed with an information specialist, including variations of the terms *artificial intelligence* and *diagnostic imaging*. The full search strategy is included in Appendix E1 (online) and Tables E1 and E2 (online). We searched the following electronic databases for English language peer-reviewed and gray literature between January 2018 and July 2020 (updated in June 2021): Ovid Medline, Ovid Embase, EBSCO Cumulative Index to Nursing and Allied Health Literature, Web of Science, Cochrane Central, Institute of Electrical and Electronics Engineers and Institution of Engineering and Technology Xplore, Association for Computing Machinery Digital Library, arXiv, medRxiv, and bioRxiv. The reference lists of all included articles were screened to identify relevant publications that were missed from our search.

We included all articles that fulfilled the following inclusion criteria: primary research studies that developed and/or validated a deep learning algorithm for fracture detection or classification in any user-independent imaging modality, English language, and human subjects. We applied the following exclusion criteria to our search: conference abstracts, letters to the editor, review articles, and studies that performed purely segmentation tasks or radiomics analysis. We excluded duplicates by using Endnote ×9, following the method described by Falconer (15). We did not place any limits on the target population, study setting, or comparator group.

Data Extraction

Titles and abstracts were screened before full-text screening. Data were extracted by using a predefined data extraction sheet. A list of excluded studies, including the reason for exclusion, was recorded in a Preferred Reporting Items for Systematic Reviews and Meta-Analyses flow diagram. Any further papers identified through reference lists underwent the same process of screening and data extraction in duplicate.

We extracted information from each study including peer-review status, study design, target condition, sample size, comparator groups, and results. Where possible, we extracted diagnostic performance information to construct contingency tables for each model and used them to calculate sensitivity and specificity. When studies included more than one contingency table, they were included in the analysis.

Statistical Analysis

We estimated the diagnostic performance of the deep learning algorithms and clinicians by carrying out a meta-analysis of studies providing contingency tables at both internal and external validation. We planned to perform a meta-analysis if at least five studies were eligible for inclusion, recommended for random-effects meta-analysis (16). We used the contingency tables to construct hierarchical summary receiver operating characteristic curves and to calculate pooled sensitivities and specificities, anticipating a high level of heterogeneity (17). We constructed a visual representation of between-study heterogeneity by using a 95% prediction region in the hierarchical summary receiver operating characteristic curves. We performed a meta-regression analysis to identify sources of between-studies heterogeneity by introducing level of bias; study and fracture type; the reference standard; peer-review status; and whether the algorithm used single or multiple radiologic views, data augmentation, or trans-

Table 2: Characteristics of Studies Developing, Internally and Externally Validating Algorithms

First Author	Year	Imaging Modality	Target Condition	View	Comparison Group	No. of Images per Set			Reference Standard	Model Output	Peer Review Status
						Training	Tuning	Testing			
Bluthgen (39)	2019	Radiography	Distal radius fracture	Concatenated AP and lateral views	Expert clinicians	524	N	300	Expert consensus	Probability of fracture and saliency map	Yes
Cheng[*] (33)	2021	Radiography	Proximal femur and pelvic fractures	AP view	Expert clinicians	NR	…	…	NR	Probability of fracture, saliency map and point annotation	Yes
Cheng[†] (38)	2019	Radiography	Proximal femur fracture	AP view	Expert clinicians	23 288	N	5822	Single expert opinion	Probability of fracture and saliency map	Yes
Choi[‡] (32)	2021	Radiography	Proximal femur fracture	AP view	None	NR	…	…	CT report	Probability of fracture and saliency map	No
Choi[§] (36)	2020	Radiography	Upper humerus fracture	AP or lateral view	Expert clinicians	1012	254	N	Expert consensus	Probability of fracture and saliency map	Yes
Lindsey[‖] (54)	2018	Radiography	Any wrist fracture	Any view	Expert and nonexpert clinicians, with and without algorithm assistance	28 341	3149	3500	Expert consensus	Binary classification and segmentation prediction	Yes
Thian (44)	2019	Radiography	Distal radius fracture	AP or lateral view	None	13 153	N	1461	Expert consensus	Saliency map	Yes
Wang (37)	2019	Radiography	Proximal femur or pelvic fracture	AP view	Expert clinicians	3087	882	441	Expert consensus	Binary classification with bounding box	No
Zhou (59)	2020	CT	Rib fractures	NR	Expert clinicians, with and without algorithm assistance	876	98	105	Expert consensus	Binary classification	Yes

Note.—AP = anteroposterior, N = no tuning set, NR = not reported.

[*] Fivefold cross-validation (n = 5204); external test set, 1888 images.

[†] External test set, 100 images.

[‡] n = 4235; external test set, 500 images.

[§] External test set 1, 258 images; external test set 2, 95 images.

[‖] External set, 1400 images.

fer learning as covariates. Statistical significance was indicated at a P value of .05. All calculations were performed by using statistical software (Stata version 14.2, Midas and Metandi modules; StataCorp) (18,19).

Quality Assessment

We assessed studies for adherence to reporting guidelines by using the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) checklist, which is a 22-item list of recommendations to aid transparent reporting of studies that develop and/or validate prediction models (20). We used a modified version of TRIPOD (Appendix E2, Table E3 [online]), as we considered that not all items on the checklist were informative for deep learning studies; for example, reporting follow-up time is irrelevant for diagnostic accuracy studies. The checklist therefore is limited in granular

Table 3: Characteristics of Studies Making Incremental Changes, or Externally Validating Algorithms

First Author	Year	Imaging Modality	Target Condition	View	Comparison Group	No. of Images per Set			Reference Standard	Model Output	Peer Review Status
						Training	Tuning	Testing			
Cheng* (34)	2020	Radiography	Proximal femur fracture	AP view	Expert and nonexpert clinicians, with and without algorithm assistance	…	…	…	Expert consensus	Probability of fracture and saliency map	Yes
Duron† (31)	2021	Radiography	Any appendicular fracture	Unspecified views	Expert and nonexpert clinicians, with and without algorithm assistance	…	…	…	Expert consensus	Binary classification and bounding box	Yes
Kitamura (56)	2019	Radiography	Ankle fracture	AP or lateral view	Comparison of multiple algorithms	1441	N	240	Expert consensus	NR	Yes
Kolanu‡ (63)	2020	CT	Vertebral compression fractures	NR	None	…	…	…	Expert consensus	NR	Yes
Uysal (25)§	2021	Radiography	Any shoulder fracture	Views not specified	Comparison of multiple algorithms	8379	N	563	NR	Binary classification	No

Note.—AP = anteroposterior, N = no tuning set NR = not reported.

* External test set, 100 images; prospective clinical trial, 632 images.

† External test set, 600 images.

‡ External validation, 1696 images.

§ External validation, 150 images.

discrimination between studies, but instead acts as a general indicator of reporting standards.

We used the Prediction Model Study Risk of Bias Assessment Tool, or PROBAST, checklist to assess papers for bias and applicability (Appendix E2, Table E4 [online]) (21). This tool uses signaling questions in four domains (participants, predictors, outcomes, and analysis) to provide both an overall and a granular assessment. We considered both the images used to develop algorithms, and the patient population or populations the models were tested on, to assess bias and applicability in the first domain. We did not include an assessment of bias or applicability for predictors. The diagnostic performance of both AI and clinicians at internal and external validation was examined separately in studies assessed at low risk of bias.

Publication Bias

We minimized the effect of publication bias by searching preprint servers and hand-searching the reference lists of included studies. We performed a formal assessment of publication bias through a regression analysis by using diagnostic log odds ratios and testing for asymmetry (22).

Results

Study Selection and Characteristics

We identified 8783 peer-reviewed studies, of which 1981 were duplicates. A further 149 studies were identified through preprint servers and citation searching. After full-text screening, 42 studies were included in the review, of which 35 were peer-reviewed publications and seven were preprint publications (Fig 1). Thirty-seven studies identified fractures on radiographs, of which 18 focused on lower limb, 15 on upper limb, and four on other fractures (Tables 1–3) (23–58). Five studies identified fractures on CT images (59–63). All studies performed their analyses with a computer, with retrospectively collected data, by using a supervised learning approach; and one study also performed a prospective nonrandomized clinical trial (34). Thirty-six studies developed and internally validated an algorithm, and nine of these studies also externally validated their algorithm (23,24,26–30,32,33,35–55,57–59,61–63). Six studies externally validated or made an incremental change to a previously developed algorithm (25,31,34,56,60,63). Twenty-three studies restricted their analysis to a single radiologic view (25,27,29–35,37,38,40–43,46–48,50–53,58).

Sixteen studies compared the performance of AI with expert clinicians, seven compared AI to experts and nonexperts, and one compared AI to nonexperts only (24,26–31,33–39,42,43,46,50,51,53–55,58,59). Six studies included clinician performance with and without AI assistance as a comparison group (29–31,34,54,59). The size of comparison groups ranged from three to 58 (median, six groups; interquartile range, 4–15). Three studies compared their algorithm against other algorithms and 16 studies did not include a comparison group (23,25,32,40,41,44–49,52,56,57,60–64).

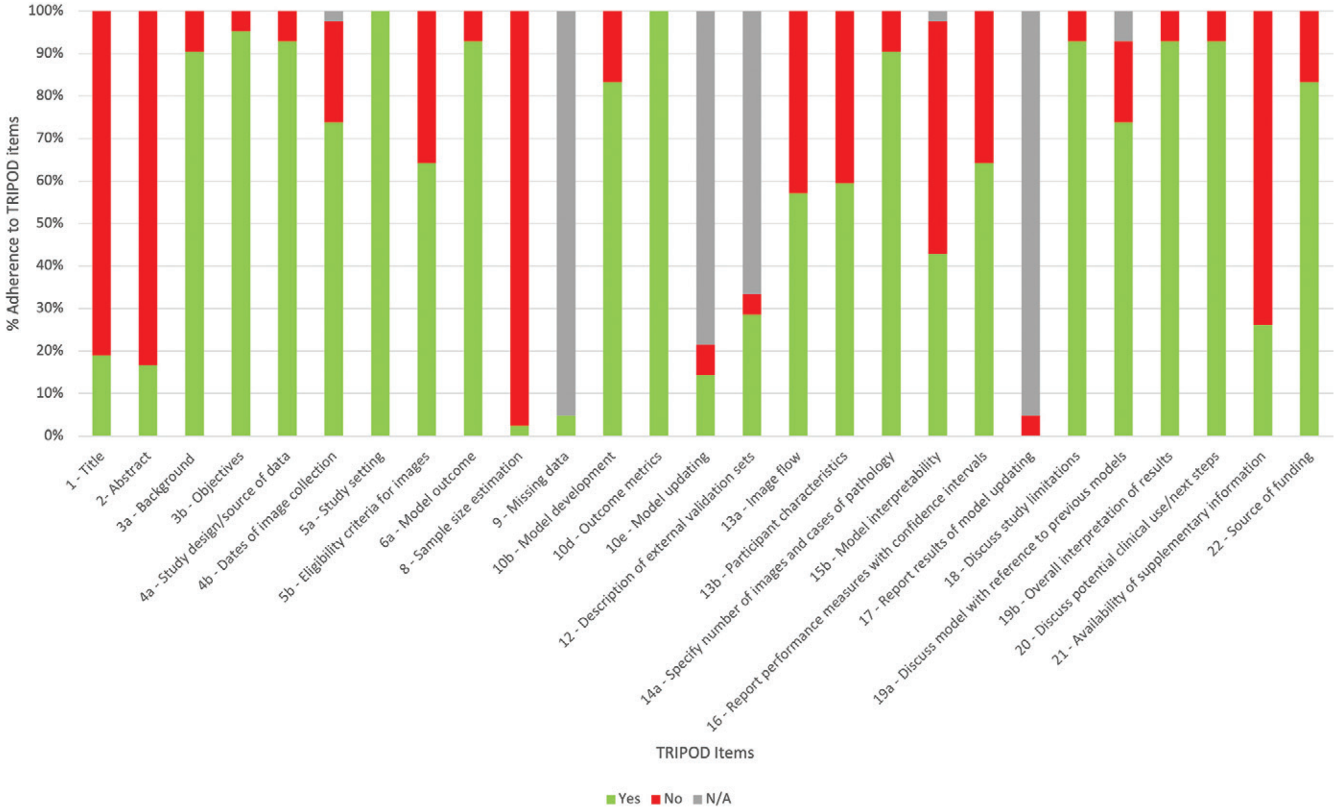

Figure 2: Summary of study adherence to Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) reporting guidelines.

To generate a reference standard for image labeling, 18 studies used expert consensus, six relied on the opinion of a single expert reader, seven used pre-existing radiologic reports or other imaging modalities, five studies used mixed methods, and one defined their reference standard as surgically confirmation fractures (23,24,26–32,34–39,41–63). Three studies did not report how their reference standard was generated (25,33,40).

Study Participants

The number of participants represented by each data set ranged widely (median, 1169; range 65–21,456; interquartile range, 425–2417; Appendix E3, Table E5 [online]). The proportion of participants with a fracture in each data set also ranged widely (median, 50%; interquartile range, 40%–64%). Seventeen studies did not include the proportion of study participants who were men or women, and 15 studies did not include information about participant age (23,25,27,34–37,41,46,54,56,57,58,60–63).

Algorithm Development and Model Output

The size of training (median, 1898; interquartile range, 784–7646), tuning (median, 739; interquartile range, 142–980) and test (median, 306; interquartile range, 233–1111) data sets at the patient level varied widely (Tables 1–3). Five of 33 (15.2%) studies that developed an algorithm did not report the size of each data set separately (24,28,32,33,40). In studies that performed external validation of an algorithm, the median size of the data set was 511 (range, 100–1696). Thirty studies used data augmentation, and 30 studies used transfer learning (23–30,32,33,35–44,46,48–59,61,62). Twenty-six studies used random split sample validation as a method of internal validation, five used stratified split sampling, and four used a resampling method (Table E6 [online]) (23–31,33,35–55,57,58,61,62).

Twenty-two studies included localization of fractures in model output to improve end-user interpretability (23,24,30–39,41,42,44,48,49,52,54,57,60,64). Metrics used to evaluate model performance varied widely, including sensitivity and specificity (38 studies); area under the receiver operating characteristic curve and Youden index (23 studies); accuracy (22 studies); positive and negative predictive values (nine studies); and F1, precision, and recall (nine studies).

Quality Assessment

Adherence to TRIPOD reporting standards was variable (Fig 2). Four items were poorly reported (<50% adherence): clarity of study title and abstract (19% and 17% adherence, respectively), sample size calculation (2.4%), discussion and attempt to improve model interpretability (43%), and a statement about supplementary code or data availability (19%).

Prediction Model Study Risk of Bias Assessment Tool, or PROBAST, led to an overall rating of 22 (52%) and 21 (50%) studies as high risk of bias and concerns regarding applicability, respectively (Fig 3). The main contributing factors to this assessment were studies that did not perform external validation, or internally validated models with small sample sizes. Fifteen (36%) studies were judged to be at high risk of bias and 18 (43%) at high concern for applicability in participant selection

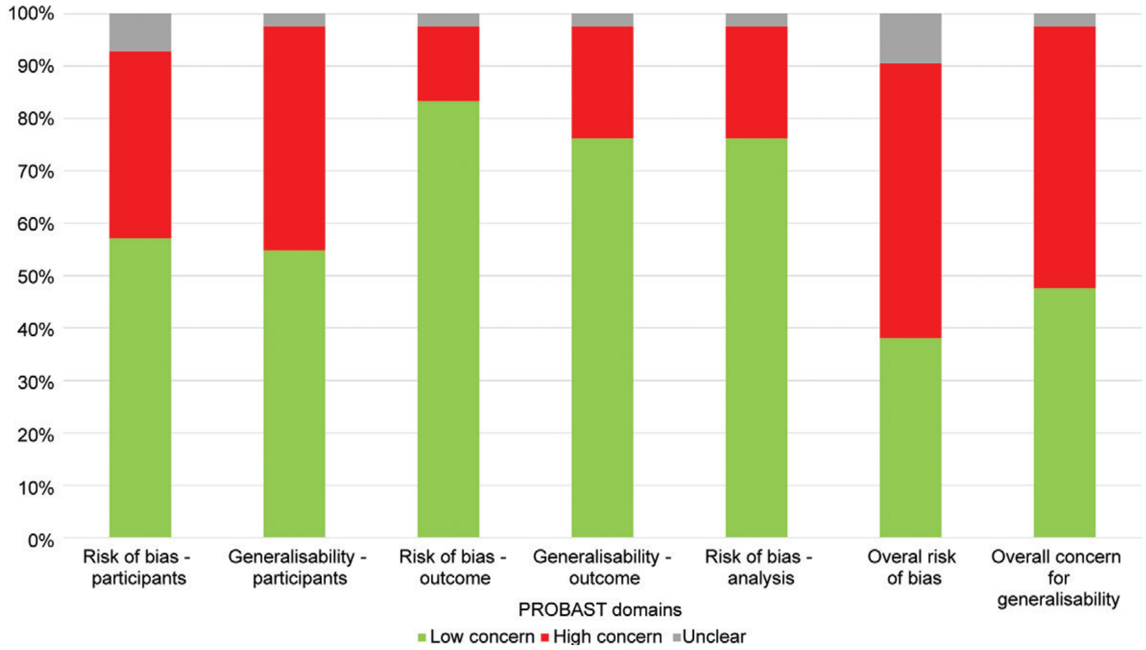

Figure 3: Summary of Prediction Model Study Risk of Bias Assessment Tool (PROBAST) risk of bias and concern about generalizability scores.

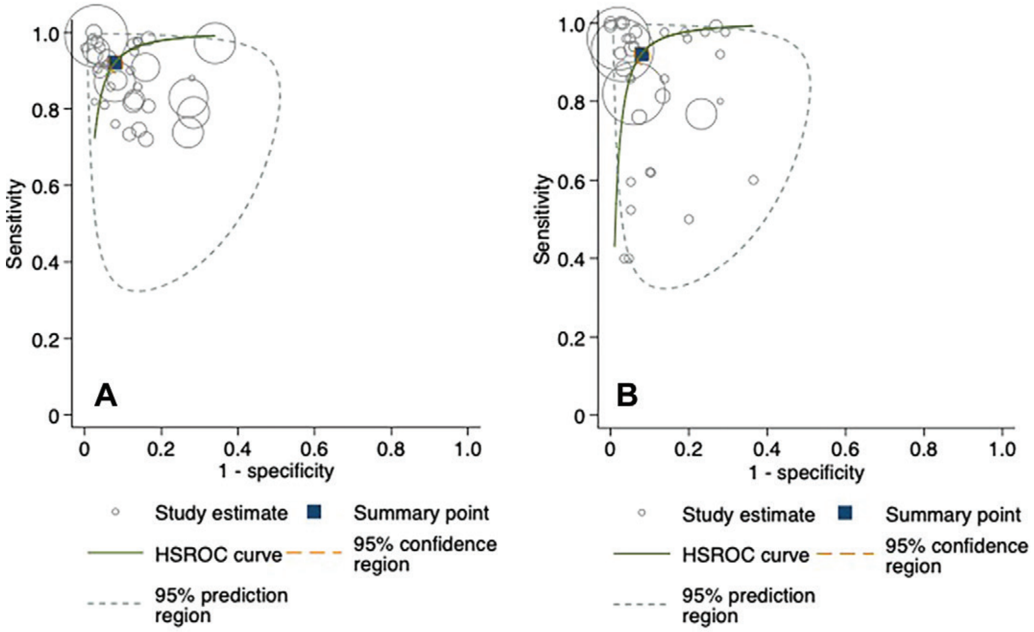

Figure 4: Hierarchical summary receiver operating characteristic (HSROC) curves for **(A)** fracture detection algorithms and **(B)** clinicians with internal validation test sets. The 95% prediction region is a visual representation of between-study heterogeneity.

because of inclusion and exclusion criteria. In general, studies were at low concern for bias (six; 14% high concern) and applicability (nine; 21% high concern) in specifying outcomes and in analysis (nine; 21% high concern).

Meta-Analysis

We extracted 115 contingency tables from 32 studies (55 061 images) that provided sufficient information to calculate contingency tables for binary fracture detection (Tables E7, E8 [online]) (23–25,27–40,42,43,45,47–49,51,53,55–58,60,61,63). Thirty-seven contingency tables from 26 studies were extracted for reported algorithm performance on internal validation, and 15 were extracted from seven studies on external validation. Thirty-six contingency tables from 12 studies were extracted for human performance on the same internal validation test sets, and 23 contingency tables from seven studies were extracted for performance on the same external validation test sets (24,27,30,31,33–39,42,43,51,53,55). Four contingency

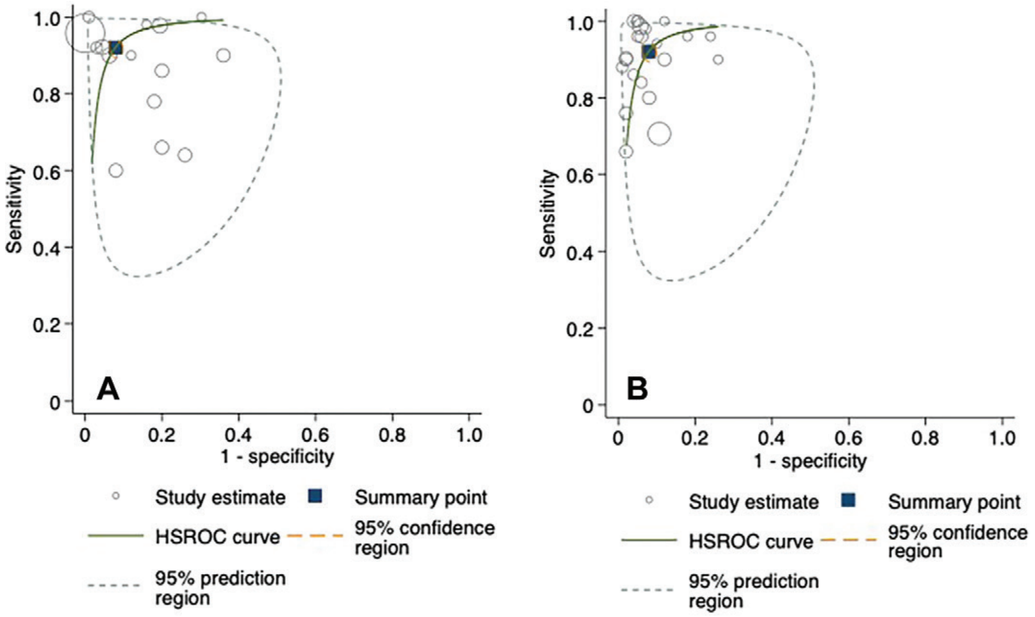

Figure 5: Hierarchical summary receiver operating characteristic (HSROC) curves for **(A)** fracture detection algorithms and **(B)** clinicians with external validation test sets. The 95% prediction region is a visual representation of between-study heterogeneity.

Table 4: Pooled Sensitivities, Specificities, and Areas Under the Curve for Artificial Intelligence Algorithms and Clinicians

Parameter	Sensitivity (%)	Specificity (%)	AUC	No. of Contingency Tables
Algorithms, internal validation, all studies	92 (88, 94)	91 (88, 93)	0.97 (0.95, 0.98)	37
Studies with low bias	90 (86, 93)	89 (85, 92)	0.95 (0.93, 0.97)	21
Clinicians, internal validation, all studies	91 (85, 95)	92 (89, 95)	0.97 (0.95, 0.98)	36
Studies with low bias	89 (76, 95)	86 (80, 90)	0.93 (0.90, 0.95)	13
Algorithms, external validation, all studies	91 (84, 85)	91 (81, 95)	0.96 (0.94, 0.98)	15
Studies with low bias	89 (76, 95)	80 (74, 85)	0.87 (0.84, 0.90)	10
Clinicians, external validation, all studies	94 (90, 96)	94 (91, 95)	0.98 (0.96, 0.99)	23
Studies with low bias	93 (87, 96)	93 (89, 95)	0.97 (0.95, 0.98)	16
Clinicians with AI assistance, all studies	97 (83, 99)	92 (88, 95)	0.95 (0.92, 0.96)	4
Studies with low bias	97 (83, 99)	92 (88, 95)	0.95 (0.92, 0.96)	4

Note.—Data in parentheses are 95% CIs. Results of all studies and studies with low bias are compared. AI = artificial intelligence, AUC = area under the receiver operating characteristic curve.

tables were extracted from four studies for human performance with AI assistance (29–31,34).

Hierarchical summary receiver operating characteristic curves from the studies evaluating AI and clinician performance on internal validation test sets are included in Figure 4. The pooled sensitivity was 92% (95% CI: 88, 94) for AI and 91% (95% CI: 85, 95) for clinicians. The pooled specificity was 91% (95% CI: 88, 93) for AI and 92% (95% CI: 89, 95) for clinicians. At external validation, the pooled sensitivity was 91% (95% CI: 84, 95) for AI and 94% (95% CI: 90, 96) for clinicians on matched test sets (Fig 5). The pooled specificity was 91% (95% CI: 82, 96) for AI and 94% (95% CI: 91, 95) for clinicians. When clinicians were provided with AI assistance, the pooled sensitivity and specificity were 97% (95% CI: 83, 99) and 92% (95% CI: 88, 95), respectively.

Meta-regression of all studies showed that lower model specificity was associated with lower risk of bias (89%; 95% CI: 87, 91; *P* , .01), use of data augmentation (92%; 95% CI: 90, 93; *P* , .01), and transfer learning (91%; 95% CI: 90, 93; *P* , .01). Higher model sensitivity was associated with algorithms focusing on lower limb fractures (95%; 95% CI: 93, 97; *P* , .01) and use of resampling methods (97%; 95% CI: 94, 100; *P* , .01). We performed a sensitivity analysis, separately evaluating studies with low risk of bias. We found that all performance metrics were lower, although only the reduction in area under the curve in studies assessing the performance of algorithms at external validation reached statistical significance (96%; 95% CI: 94, 98; *P* , .01; Table 4, Fig 6). We report findings of sensitivity analyses for other covariates in Figure E1, Appendix E4, and Tables E9–E13 (online).

Publication Bias

We assessed publication bias by using a regression analysis to quantify funnel plot asymmetry (Fig E2 [online]) (22). We

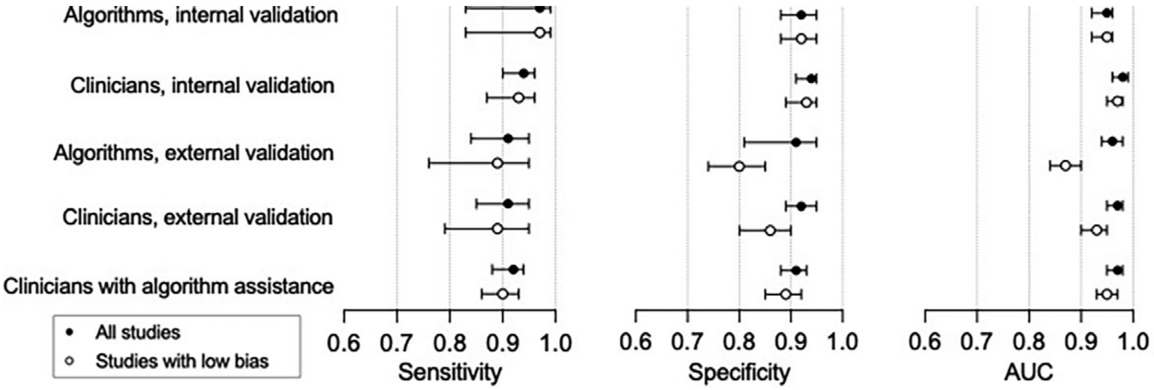

Figure 6: Summary of pooled sensitivity, specificity, and area under the curve (AUC) of algorithms and clinicians comparing all studies versus low-bias studies with 95% CIs.

found that the slope coefficient was −5.4 (95% CI: −13.7, 2.77; *P* = .19), suggesting a low risk of publication bias.

Discussion

An increasing number of studies are investigating the potential for artificial intelligence (AI) as a diagnostic adjunct in fracture diagnosis. We performed a systematic review of the methods, results, reporting standards, and quality of studies in assessing deep learning in fracture detection tasks. We performed a meta-analysis of diagnostic performance, grouped into internal and external validation results, and compared with clinician performance. Our review highlighted four principal findings. First, AI had high reported diagnostic accuracy, with a pooled sensitivity of 91% (95% CI: 84, 85) and specificity of 91% (95% CI: 81, 95). Second, AI and clinicians had comparable performance (pooled sensitivity, 94% [95% CI: 90, 96]; and specificity, 94% [95% CI: 91, 95]) at external validation. The addition of AI assistance improved clinician performance further (pooled sensitivity, 97% [95% CI: 83, 99]; and specificity, 92% [95% CI: 88, 95]), and one study found that clinicians reached a diagnosis in a shorter time with AI assistance (29–31, 34). Third, there were significant flaws in study methods that may limit the real-world applicability of study findings. For example, it is likely that clinician performance was underestimated: Only one study provided clinicians with background clinical information. Half of the studies that had a clinician comparison group used small groups (ie, less than five) at high risk of interrater variation. All studies performed experiments on a computer or via computer simulation, and only one evaluated human-algorithm performance in a prospective clinical trial. Fourth, there was high heterogeneity across studies, partly attributable to variations in study methods. Heterogeneity in sensitivity and specificity was higher when methodologic choices, such as internal validation methods or reference standards, were used. There was a wide range of study sample size, but only one study (63) performed a sample size calculation.

Previous narrative reviews have reported a wide range of AI accuracy (11–13). However, the use of accuracy as an outcome metric in image classification tasks can be misleading (65). For example, in a data set consisting of 82% fracture and 18% unfractured radiographs, an AI that always predicts a fracture will have a reported accuracy of 82%, despite being deeply flawed (30). A meta-analysis of nine studies by Yang et al (14) reported a pooled sensitivity and specificity of 87% (95% CI: 78, 93) and 91% (95% CI: 85, 95), respectively. This is consistent with the findings of our meta-analysis of 32 studies. We provided further granularity of results, reporting pooled sensitivity and specificity separately for internal (sensitivity, 92% [95% CI: 88, 94]; and specificity, 91% [95% CI: 88, 93]) and external (sensitivity, 91% [95% CI: 84, 95]; and specificity, 91% [95% CI: 81, 95]) validation.

Our study had limitations. First, we only included studies in the English language that were published after 2018, excluding other potentially eligible studies. Second, we were only able to extract contingency tables from 32 studies. Third, many studies had methodologic flaws and half were classified as high concern for bias and applicability, limiting the conclusions that could be drawn from the meta-analysis because studies with high risk of bias consistently overestimated algorithm performance. Fourth, although adherence to TRIPOD items was generally fair, many manuscripts omitted vital information such as the size of training, tuning, and test sets.

The results from this meta-analysis cautiously suggest that AI is noninferior to clinicians in terms of diagnostic performance in fracture detection, showing promise as a useful diagnostic tool. Many studies have limited real-world applicability because of flawed methods or unrepresentative data sets. Future research must prioritize pragmatic algorithm development. For example, imaging views may be concatenated, and databases should mirror the target population (eg, in fracture prevalence, and age and sex of patients). It is crucial that studies include an objective assessment of sample size adequacy as a guide to readers (66). Data and code sharing across centers may spread the burden of generating large and precisely labeled data sets, and this is encouraged to improve research reproducibility and transparency (67,68). Transparency of study methods and clear presentation of results is necessary for accurate critical appraisal. Machine learning extensions to TRIPOD, or TRIPOD-ML, and Standards for Reporting of Diagnostic Accuracy Studies, or STARD-AI, guidelines are currently being developed and may improve conduct and reporting of deep learning studies (69–71).

Future research should seek to externally validate algorithms in prospective clinical settings and provide a fair comparison with relevant clinicians: for example, providing clinicians with routine clinical detail. External validation and evaluation of algorithms in prospective randomized clinical trials is a necessary next step toward clinical deployment. Current artificial intelligence (AI) is designed as a diagnostic adjunct and may improve workflow through screening or prioritizing images on worklists and highlighting regions of interest for a reporting radiologist. AI may also improve diagnostic certainty through acting as a "second reader" for clinicians or as an interim report prior to radiologist interpretation. However, it is not a replacement for the clinical workflow, and clinicians must understand AI performance and exercise judgement in interpreting algorithm output. We advocate for transparent reporting of study methods and results as crucial to AI integration. By addressing these areas for development, deep learning has potential to streamline fracture diagnosis in a way that is safe and sustainable for patients and health care systems.

Acknowledgment: We thank Eli Harriss, BA, MS, Bodleian Libraries Outreach Librarian for the Bodleian Health Care Libraries, who formulated the search strategies and ran the database searches.

Author contributions: Guarantors of integrity of entire study, **R.Y.L.K., D.F.**; 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, **R.Y.L.K., C.H., T.A.C., B.J., A.F., D.C., D.F.**; statistical analysis, **R.Y.L.K., D.C., G.S.C.**; and manuscript editing, **R.Y.L.K., C.H., B.J., A.F., D.C., M.S., G.S.C., D.F.**

Data sharing: All data generated or analyzed during the study are included in the published paper.

Disclosures of conflicts of interest: **R.Y.L.K.** No relevant relationships. **C.H.** No relevant relationships. **T.A.C.** No relevant relationships. **B.J.** No relevant relationships. **A.F.** No relevant relationships. **D.C.** No relevant relationships. **M.S.** No relevant relationships. **G.S.C.** No relevant relationships. **D.F.** Chair, British Society for Surgery of the Hand Research Committee; member, British Association of Plastic, Reconstructive, and Aesthetic Surgeons Research Committee; member, British Lymphology Society Research Committee; chair, Scientific Advisory Committee Restore Research; Trustee, British Dupuytren Society.

References

1. Bergh C, Wennergren D, Möller M, Brisby H. Fracture incidence in adults in relation to age and gender: A study of 27,169 fractures in the Swedish Fracture Register in a well-defined catchment area. PLoS One 2020;15(12):e0244291.
2. Amin S, Achenbach SJ, Atkinson EJ, Khosla S, Melton LJ 3rd. Trends in fracture incidence: a population-based study over 20 years. J Bone Miner Res 2014;29(3):581–589.
3. Curtis EM, van der Velde R, Moon RJ, et al. Epidemiology of fractures in the United Kingdom 1988-2012: Variation with age, sex, geography, ethnicity and socioeconomic status. Bone 2016;87:19–26.
4. UK NHS Annual Report. Hospital accident & emergency activity 2019-20. https://digital.nhs.uk/data-and-information/publications/statistical/hospital-accident–emergency-activity/2019-20. Accessed December 21, 2021.
5. Wei CJ, Tsai WC, Tiu CM, Wu HT, Chiou HJ, Chang CY. Systematic analysis of missed extremity fractures in emergency radiology. Acta Radiol 2006;47(7):710–717.
6. Williams SM, Connelly DJ, Wadsworth S, Wilson DJ. Radiological review of accident and emergency radiographs: a 1-year audit. Clin Radiol 2000;55(11):861–865.
7. Hallas P, Ellingsen T. Errors in fracture diagnoses in the emergency department--characteristics of patients and diurnal variation. BMC Emerg Med 2006;6(1):4.
8. Zha N, Patlas MN, Duszak R Jr. Radiologist burnout is not just isolated to the united states: Perspectives from Canada. J Am Coll Radiol 2019;16(1):121–123.
9. Bender CE, Bansal S, Wolfman D, Parikh JR. 2018 ACR commission on human resources workforce survey. J Am Coll Radiol 2019;16(4 Pt A):508–512.
10. McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. Nature 2020;577(7788):89–94 [Published correction appears in Nature 2020;586(7829):E19.].
11. Smets J, Shevroja E, Hügle T, Leslie WD, Hans D. Machine learning solutions for Osteoporosis—A review. J Bone Miner Res 2021;36(5):833–851.
12. Langerhuizen DWG, Janssen SJ, Mallee WH, et al. What are the applications and limitations of artificial intelligence for fracture detection and classification in orthopaedic trauma imaging? A systematic review. Clin Orthop Relat Res 2019;477(11):2482–2491.
13. Kalmet PHS, Sanduleanu S, Primakov S, et al. Deep learning in fracture detection: a narrative review. Acta Orthop 2020;91(2):215–220.
14. Yang S, Yin B, Cao W, Feng C, Fan G, He S. Diagnostic accuracy of deep learning in orthopaedic fractures: A systematic review and meta-analysis. Clin Radiol 2020;75(9):713.e17–713.e28.
15. Falconer J. Removing duplicates from an EndNote library/2021. http://blogs.lshtm.ac.uk/library/2018/12/07/removing-duplicates-from-an-endnote-library/. Accessed May 6, 2021.
16. Jackson D, Turner R. Power analysis for random-effects meta-analysis. Res Synth Methods 2017;8(3):290–302.
17. Macaskill P, Gatsonis C, Deeks J, Harbord R, Takwoingi Y. Cochrane handbook for systematic reviews of diagnostic test accuracy. Version 0.9.0. London, England: The Cochrane Collaboration, 2010; 83.
18. Harbord RM, Whiting P. Metandi: Meta-analysis of diagnostic accuracy using hierarchical logistic regression. Stata J 2009;9(2):211–229.
19. Dwamena B. MIDAS: Stata module for meta-analytical integration of diagnostic test accuracy studies. https://ideas.repec.org/c/boc/bocode/s456880.html. Published 2009. Accessed January 2, 2022.
20. Collins GS, Reitsma JB, Altman DG, Moons KG; TRIPOD Group. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. The TRIPOD Group. Circulation 2015;131(2):211–219.
21. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med 2019;170(1):51–58.
22. Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. J Clin Epidemiol 2005;58(9):882–893.
23. Yoon AP, Lee YL, Kane RL, Kuo CF, Lin C, Chung KC. Development and validation of a deep learning model using convolutional neural networks to identify scaphoid fractures in radiographs. JAMA Netw Open 2021;4(5):e216096.
24. Raisuddin AM, Vaattovaara E, Nevalainen M, et al. Critical evaluation of deep neural networks for wrist fracture detection. Sci Rep 2021;11(1):6006.
25. Uysal F, Hardalaç F, Peker O, Tolunay T, Tokgöz N. Classification of shoulder X-ray images with deep learning ensemble models. Appl Sci (Basel) 2021;11(6):2723.
26. Yamada Y, Maki S, Kishida S, et al. Automated classification of hip fractures using deep convolutional neural networks with orthopedic surgeon-level accuracy: ensemble decision-making with antero-posterior and lateral radiographs. Acta Orthop 2020;91(6):699–704.
27. Ozkaya E, Topal FE, Bulut T, Gursoy M, Ozuysal M, Karakaya Z. Evaluation of an artificial intelligence system for diagnosing scaphoid fracture on direct radiography. Eur J Trauma Emerg Surg 2020. 10.1007/s00068-020-01468-0. Published online August 30, 2020.
28. Murata K, Endo K, Aihara T, et al. Artificial intelligence for the detection of vertebral fractures on plain spinal radiography. Sci Rep 2020;10(1):20031.
29. Mawatari T, Hayashida Y, Katsuragawa S, et al. The effect of deep convolutional neural networks on radiologists' performance in the detection of hip fractures on digital pelvic radiographs. Eur J Radiol 2020;130:109188.
30. Krogue JD, Cheng KV, Hwang KM, et al. Automatic hip fracture identification and functional subclassification with deep learning. Radiol Artif Intell 2020;2(2):e190023.

31. Duron L, Ducarouge A, Gillibert A, et al. Assessment of an AI aid in detection of adult appendicular skeletal fractures by emergency physicians and radiologists: A multicenter cross-sectional diagnostic study. Radiology 2021;300(1):120–129.
32. Choi J, Hui JZ, Spain D, Su YS, Cheng CT, Liao CH. Practical computer vision application to detect hip fractures on pelvic X-rays: a bi-institutional study. Trauma Surg Acute Care Open 2021;6(1):e000705.
33. Cheng CT, Wang Y, Chen HW, et al. A scalable physician-level deep learning algorithm detects universal trauma on pelvic radiographs. Nat Commun 2021;12(1):1066.
34. Cheng CT, Chen CC, Cheng FJ, et al. A human-algorithm integration system for hip fracture detection on plain radiography: System development and validation study. JMIR Med Inform 2020;8(11):e19416.
35. Chen HY, Hsu BW, Yin YK, et al. Application of deep learning algorithm to detect and visualize vertebral fractures on plain frontal radiographs. PLoS One 2021;16(1):e0245992.
36. Choi JW, Cho YJ, Lee S, et al. Using a dual-input convolutional neural network for automated detection of pediatric supracondylar fracture on conventional radiography. Invest Radiol 2020;55(2):101–110.
37. Wang Y, Lu L, Cheng C, et al. Weakly supervised universal fracture detection in pelvic x-rays. In: Shen D, Liu T, Peters TM, et al, eds. Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. MICCAI 2019. Lecture Notes in Computer Science, vol 11769. Cham, Switzerland: Springer, 2019; 459–467.
38. 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.
39. 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.
40. Beyaz S, Açıcı K, Sümer E. Femoral neck fracture detection in X-ray images using deep learning and genetic algorithm approaches. Jt Dis Relat Surg 2020;31(2):175–183.
41. Yahalomi E, Chernofsky M, Werman M. Detection of distal radius fractures trained by a small set of X-ray images and faster R-CNN. arXiv preprint arXiv:1812.09025. https://arxiv.org/abs/1812.09025. Posted December 21, 2018. Accessed May 6, 2021.
42. Yu JS, Yu SM, Erdal BS, et al. Detection and localisation of hip fractures on anteroposterior radiographs with artificial intelligence: proof of concept. Clin Radiol 2020;75(3):237.e1–237.e9.
43. Urakawa T, Tanaka Y, Goto S, Matsuzawa H, Watanabe K, Endo N. Detecting intertrochanteric hip fractures with orthopedist-level accuracy using a deep convolutional neural network. Skeletal Radiol 2019;48(2):239–244.
44. Thian YL, Li Y, Jagmohan P, Sia D, Chan VEY, Tan RT. Convolutional neural networks for automated fracture detection and localization on wrist radiographs. Radiol Artif Intell 2019;1(1):e180001.
45. Rayan JC, Reddy N, Kan JH, Zhang W, Annapragada A. Binomial classification of pediatric elbow fractures using a deep learning multiview approach emulating radiologist decision making. Radiol Artif Intell 2019;1(1):e180015.
46. 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.
47. Mehta SD, Sebro R. Computer-aided detection of incidental lumbar spine fractures from routine dual-energy X-ray absorptiometry (DEXA) studies using a support vector machine (SVM) classifier. J Digit Imaging 2020;33(1):204–210.
48. Mutasa S, Varada S, Goel A, Wong TT, Rasiej MJ. Advanced deep learning techniques applied to automated femoral neck fracture detection and classification. J Digit Imaging 2020;33(5):1209–1217.
49. Starosolski ZA, Kan H, Annapragada AV. CNN-based radiographic acute tibial fracture detection in the setting of open growth plates. bioRxiv preprint bioRxiv:506154. https://www.biorxiv.org/content/10.1101/506154. Posted January 3, 2019. Accessed May 6, 2021.
50. Jiménez-Sánchez A, Kazi A, Albarqouni S, et al. Precise proximal femur fracture classification for interactive training and surgical planning. Int J CARS 2020;15(5):847–857.
51. Gan K, Xu D, Lin Y, et al. Artificial intelligence detection of distal radius fractures: a comparison between the convolutional neural network and professional assessments. Acta Orthop 2019;90(4):394–400.
52. Derkatch S, Kirby C, Kimelman D, Jozani MJ, Davidson JM, Leslie WD. Identification of vertebral fractures by convolutional neural networks to predict nonvertebral and hip fractures: A registry-based cohort study of dual X-ray absorptiometry. Radiology 2019;293(2):405–411.
53. 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.
54. 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.
55. Langerhuizen DWG, Bulstra AEJ, Janssen SJ, et al. Is deep learning on par with human observers for detection of radiographically visible and occult fractures of the scaphoid? Clin Orthop Relat Res 2020;478(11):2653–2659.
56. 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.
57. Grauhan NF, Niehues SM, Gaudin RA, et al. Deep learning for accurately recognizing common causes of shoulder pain on radiographs. Skeletal Radiol 2022;51(2):355–362.
58. Kim DH, MacKinnon T. Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks. Clin Radiol 2018;73(5):439–445.
59. Zhou QQ, Wang J, Tang W, et al. Automatic detection and classification of rib fractures on thoracic CT using convolutional neural network: Accuracy and feasibility. Korean J Radiol 2020;21(7):869–879.
60. Weikert T, Noordtzij LA, Bremerich J, et al. Assessment of a deep learning algorithm for the detection of rib fractures on whole-body trauma computed tomography. Korean J Radiol 2020;21(7):891–899.
61. Raghavendra U, Bhat NS, Gudigar A, Acharya UR. Automated system for the detection of thoracolumbar fractures using a CNN architecture. Future Gener Comput Syst 2018;85:184–189.
62. Pranata YD, Wang KC, Wang JC, et al. Deep learning and SURF for automated classification and detection of calcaneus fractures in CT images. Comput Methods Programs Biomed 2019;171:27–37.
63. Kolanu N, Silverstone E, Pham H, et al. Utility of computer-aided vertebral fracture detection software. JOURNAL 2020;31(Suppl 1):S179.
64. Sato Y, Takegami Y, Asamoto T, et al. A computer-aided diagnosis system using artificial intelligence for hip fractures -multi-institutional joint development research-. arXiv preprint arXiv:2003.12443. https://arxiv.org/abs/2003.12443. Posted March 11, 2020. Accessed May 6, 2021.
65. Kuo RYL, Harrison CJ, Jones BE, Geoghegan L, Furniss D. Perspectives: A surgeon's guide to machine learning. Int J Surg 2021;94:106133.
66. Balki I, Amirabadi A, Levman J, et al. Sample-size determination methodologies for machine learning in medical imaging research: A systematic review. Can Assoc Radiol J 2019;70(4):344–353.
67. Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AK; SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI Extension. BMJ 2020;370:m3164.
68. Rivera SC, Liu X, Chan AW, Denniston AK, Calvert MJ; SPIRIT-AI and CONSORT-AI Working Group. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI Extension. BMJ 2020;370:m3210.
69. Turner L, Shamseer L, Altman DG, Schulz KF, Moher D. Does use of the CONSORT Statement impact the completeness of reporting of randomised controlled trials published in medical journals? A Cochrane review. Syst Rev 2012;1(1):60.
70. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. Lancet 2019;393(10181):1577–1579.
71. Sounderajah V, Ashrafian H, Aggarwal R, et al. Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: The STARD-AI Steering Group. Nat Med 2020;26(6):807–808.