Clin Orthop Relat Res (2019) 477:2482-2491
DOI 10.1097/CORR.0000000000000848

Clinical Orthopaedics and Related Research®
A Publication of The Association of Bone and Joint Surgeons®

Systematic Review

What Are the Applications and Limitations of Artificial Intelligence for Fracture Detection and Classification in Orthopaedic Trauma Imaging? A Systematic Review

David W. G. Langerhuizen MD, Stein J. Janssen MD, PhD, Wouter H. Mallee MD, Michel P.J. van den Bekerom MD, PhD, David Ring MD, PhD, Gino M. M. J. Kerkhoffs MD, PhD, Ruurd L. Jaarsma MD, PhD, FRACS, Job N. Doornberg MD, PhD

Received: 11 February 2019 / Accepted: 14 May 2019 / Published online: 21 June 2019

Abstract

Background Artificial-intelligence algorithms derive rules and patterns from large amounts of data to calculate the probabilities of various outcomes using new sets of similar data. In medicine, artificial intelligence (AI) has been applied primarily to image-recognition diagnostic tasks and evaluating the probabilities of particular

One of the authors (DWGL) certifies that he received an amount less than USD 10,000 from the Michael van Vloten Foundation (Rotterdam, the Netherlands), an amount less than USD 10,000 from the Anna Foundation (Oegstgeest, the Netherlands), and an amount less than USD 10,000 from the Traumaplatform Foundation (Amsterdam, the Netherlands). One of the authors (DR) certifies that he received payments in an amount of less than USD 10,000 in royalties from Skeletal Dynamics (Miami, FL, USA) and payments in an amount of less than USD 10,000 in personal fees from Wright Medical (Memphis, TN, USA), personal fees as Deputy Editor for *Clinical Orthopaedics and Related Research®*, personal fees from universities and hospitals, and personal fees from lawyers outside the submitted work. One of the authors (JND) certifies that he has received an unrestricted Postdoc Research Grant from the Marti-Keuning-Eckhardt Foundation.
Each author certifies that his institution waived approval for the reporting of this investigation and that all investigations were conducted in conformity with ethical principles of research.
Clinical Orthopaedics and Related Research® neither advocates nor endorses the use of any treatment, drug, or device. Readers are encouraged to always seek additional information, including FDA approval status, of any drug or device before clinical use.
This study was performed at Flinders Medical Centre, Adelaide, Australia and the Amsterdam University Medical Centre, Amsterdam, the Netherlands.

D. W. G. Langerhuizen, R. L. Jaarsma, J. N. Doornberg, Flinders University, Department of Orthopaedic and Trauma Surgery, Flinders Medical Centre, Adelaide, Australia

S. J. Janssen, Department of Orthopaedic Surgery, Amphia Hospital, Breda, the Netherlands

W. H. Mallee, G. M. M. J. Kerkhoffs, Department of Orthopaedic Surgery, Amsterdam Movement Sciences, Amsterdam University Medical Centre, Amsterdam, the Netherlands

M. P. J. van den Bekerom, Department of Orthopaedic Surgery, Onze Lieve Vrouwe Gasthuis, Amsterdam, the Netherlands

D. Ring, Department of Surgery and Perioperative Care, Dell Medical School, the University of Texas at Austin, Austin, TX, USA

D. W. G. Langerhuizen (✉), Department of Orthopaedic and Trauma Surgery, Flinders Medical Centre, Flinders University, Bedford Park, Level 5, Room 5A 153, Bedford Park 5042 South Australia, Email: david.langerhuizen@gmail.com

All ICMJE Conflict of Interest Forms for authors and *Clinical Orthopaedics and Related Research®* editors and board members are on file with the publication and can be viewed on request.

Wolters Kluwer

outcomes after treatment. However, the performance and limitations of AI in the automated detection and classification of fractures has not been examined comprehensively.

Question/purposes In this systematic review, we asked (1) What is the proportion of correctly detected or classified fractures and the area under the receiving operating characteristic (AUC) curve of AI fracture detection and classification models? (2) What is the performance of AI in this setting compared with the performance of human examiners?

Methods The PubMed, Embase, and Cochrane databases were systematically searched from the start of each respective database until September 6, 2018, using terms related to "fracture", "artificial intelligence", and "detection, prediction, or evaluation." Of 1221 identified studies, we retained 10 studies: eight studies involved fracture detection (ankle, hand, hip, spine, wrist, and ulna), one addressed fracture classification (diaphyseal femur), and one addressed both fracture detection and classification (proximal humerus). We registered the review before data collection (PROSPERO: CRD42018110167) and used the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA). We reported the range of the accuracy and AUC for the performance of the predicted fracture detection and/or classification task. An AUC of 1.0 would indicate perfect prediction, whereas 0.5 would indicate a prediction is no better than a flip-of-a-coin. We conducted quality assessment using a seven-item checklist based on a modified methodologic index for nonrandomized studies instrument (MINORS).

Results For fracture detection, the AUC in five studies reflected near perfect prediction (range, 0.95-1.0) , and the accuracy in seven studies ranged from 83% to 98%. For fracture classification, the AUC was 0.94 in one study, and the accuracy in two studies ranged from 77% to 90%. In two studies AI outperformed human examiners for detecting and classifying hip and proximal humerus fractures, and one study showed equivalent performance for detecting wrist, hand and ankle fractures.

Conclusions Preliminary experience with fracture detection and classification using AI shows promising performance. AI may enhance processing and communicating probabilistic tasks in medicine, including orthopaedic surgery. At present, inadequate reference standard assignments to train and test AI is the biggest hurdle before integration into clinical workflow. The next step will be to apply AI to more challenging diagnostic and therapeutic scenarios when there is absence of certitude. Future studies should also seek to address legal regulation and better determine feasibility of implementation in clinical practice.

Level of Evidence Level II, diagnostic study.

Introduction

In 1959, Arthur Samuel defined artificial intelligence (AI) as a field of study that gives a computer the ability to learn without needing to be reprogrammed [26]. In layman's terms, AI algorithms are developed to derive rules and patterns from large amounts of data to calculate the probabilities of various outcomes with new sets of similar data (Fig. 1). For instance, Netflix uses AI algorithms to analyze the viewing preferences of millions of people and determine what a viewer is likely to enjoy based on prior viewing behavior. Computers are programmed to continuously update probabilities of a person liking a given television show based on a combination of new all-user data and individual viewing choices.

The initial applications of AI in medicine have focused largely on image-recognition diagnostic tasks such as detecting retinopathy in diabetic people via photographs of the retinal fundus, detecting mammographic lesions, and recognizing skin cancer [6, 10, 15]. AI algorithms that address treatment probabilities—such as decision-support tools to assist orthopaedic oncologists in predicting survival and mortality—have also been developed but are not yet widely used in clinical practice [13, 29].

AI might be useful to aid the diagnostic aspects of fracture care. For example, AI applications might improve the diagnosis of true fractures among suspected fractures of the scaphoid or hip, detect key fracture characteristics that might alter prognosis and treatment, or help detect less severe fractures that are often overlooked during a secondary evaluation after complex trauma [11, 24]. The key applications of AI will help address the shortcomings of human intelligence that make us susceptible to the magician's sleight of hand and, likewise, to overlook important details in distracting circumstances. In clinical practice, both the routine and complex can be distractions.

We aggregated data from published studies using AI for fracture detection and classification to address the following questions: (1) What is the proportion of correctly detected or classified fractures and the area under the receiving operating characteristic (AUC) curve of AI fracture detection and classification models? (2) What is the performance of AI in this setting compared with the performance of human examiners?

Materials and Methods

Article Selection, Quality Assessment, and Data Extraction

We performed a systematic search according to the PRISMA statement [20] using the PubMed, Embase, and Cochrane libraries for studies from the start of each

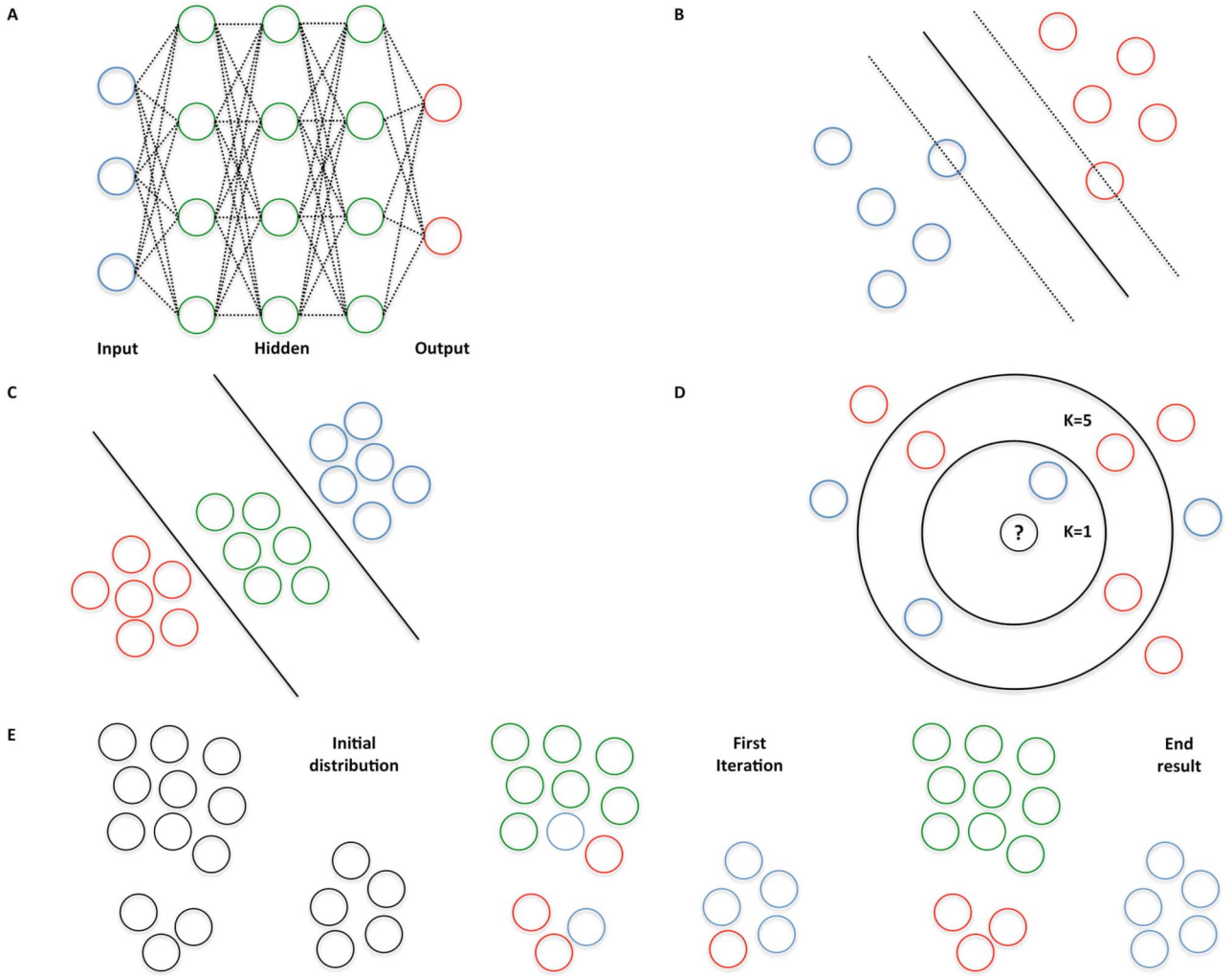

Fig. 1 A-E Two common AI techniques exist. Supervised learning applies to iterative training of an algorithm with a dataset consisting of input features with ground truth labels. For example, radiographs of the wrist are provided as input features labelled fracture and no fracture. By providing new wrist radiographs without a label, the algorithm learns to make a prediction of both classes on its own. Unsupervised learning applies to data exposure without ground truth labels. During the training phase, the algorithm attempts to find labels that best organize the data ("clustering"). Generally, unsupervised learning requires more computational power and larger datasets, and its performance is more challenging to evaluate. Therefore, supervised algorithms are often used in medical applications. **(A)** Neural networks are based on interconnected neurons in the human brain. The blue dots represent input features, whereas the red dots are the output of the algorithm. The green dots mathematically weigh the input features to predict an output. **(B)** A support vector machine is used to define an optimal separating "hyperplane" to maximize the distance from the closest points of two classes. **(C)** A linear discriminant analysis is a linear classification technique to distinguish among three or more classes. **(D)** K-nearest neighbors classify an input feature by a majority vote of its K-closest neighbors. For instance, the unknown dot will be assigned blue if K = 1 (inner circle), whereas the unknown dot will be assigned red if K = 5 (outer circle). **(E)** K-means groups objects based on their characteristics by iteratively aggregating clusters to centroids by minimizing the distance to the middle point of the cluster. For example, three clusters are aggregated (K = 3): green, red, and blue dots.

respective database until September 6, 2018. Our review protocol was registered on PROSPERO (CRD42018110167) before data collection. A professional medical librarian helped us build the search syntax using the following keywords in the title and abstract: (orthopedics OR orthopedic procedures OR traumatology OR fracture* OR skeletal fixation* OR (trauma* AND orthop*)) AND (artificial intelligence OR neural network* or deep learning OR machine learning OR machine intelligence) AND (predict* OR predictive value of test OR score OR scores OR scoring system OR scoring systems OR observ* OR observer variation OR detect* or evaluat*

OR analy* OR assess* OR measure*) (see Appendix; Supplemental Digital Content 1, http://links.lww.com/CORR/A194).

Two reviewers (DWGL, SJJ) independently screened the titles and abstracts, and if a study was considered eligible, they together screened the full-text article using predefined criteria to reach agreement. A third reviewer was not deemed necessary as a high level of consistency during the screening and inclusion process was achieved. Articles met the inclusion criteria if they addressed one or more AI models (a mathematical computing algorithm trained with "big data" to autonomously assign labels to unseen data) for detecting and/or classifying fractures on any radiologic imaging modality. We did not restrict the radiologic imaging modality to detect and/or classify fractures. We excluded studies in which patients were not in an orthopaedic trauma setting, studies evaluating robot-assisted surgery techniques, studies with mixed cohorts without clear subgroup reporting, review articles, letters to the editor, conference abstracts, technique papers, animal and cadaveric studies, and studies not published in English.

The database search yielded 1221 citations, and after removing duplicate articles, we screened 1044 potentially eligible records (Fig. 2). Twenty-eight studies were selected for full-text screening, of which eight remained. However, two additional eligible studies were identified through verbal communication in our network and meeting proceedings, but did not appear in our structured systematic searches [8, 17]. We did not identify new eligible studies through screening the reference lists of the included studies.

Two reviewers (DWGL, SJJ) independently appraised the quality of all included studies. The Newcastle-Ottawa Scale and methodologic index for nonrandomized studies (MINORS) instruments are commonly used for cohort or case-control studies [27, 32]. However, there is no risk of bias assessment tool that is suitable for diagnostic studies. Therefore, we decided to conduct quality assessment using a modified seven-item checklist based on the MINORS criteria, including disclosure, study aim, input feature, determination of ground truth labels, dataset distribution, performance metric, and explanation of the used AI model. Standardized forms were used to extract and record data using an electronic database (Microsoft Excel Version 16.21; Microsoft Inc, Redmond, WA, USA). A consensus meeting between both observers (DWGL, SJJ) was held to overcome disagreements regarding article selection, quality assessment, and data extraction.

Outcome Measures

Our primary study outcome was the proportion of correctly detected or classified fractures and nonfractures to the total number of patients and the area under the receiving operating characteristic (AUC) curve of AI models. A total of 10 studies met inclusion criteria and were used to answer this research question. Our secondary outcome was the performance of AI in this setting compared with the performance of human examiners. Three studies met inclusion criteria and were used to answer this research question.

The following data were obtained from each study: year of publication, input feature (radiologic imaging modality), projection when plain radiography was used as a radiologic imaging modality (for example, AP, oblique, or lateral views) [2, 3, 5, 8, 14, 17, 23, 30], size of the dataset,

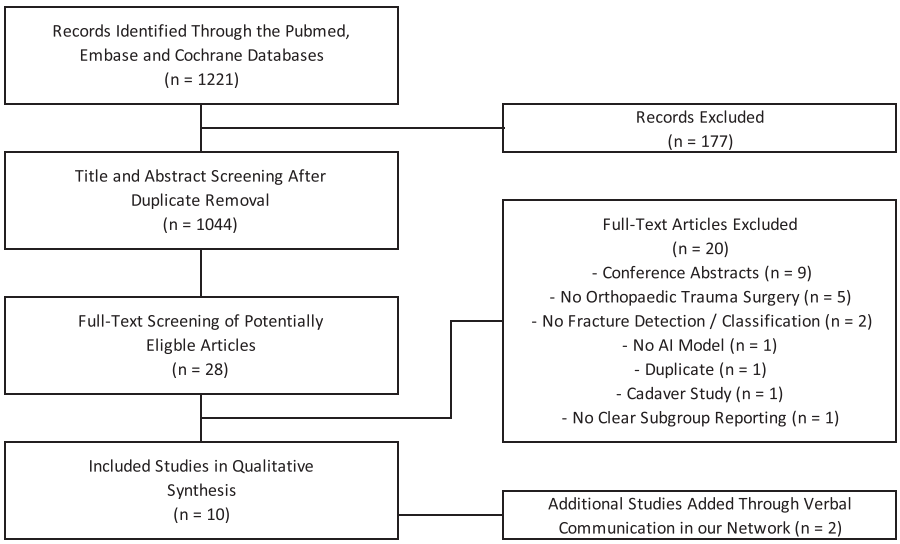

Fig. 2 This flowchart depicts the study selection during screening and inclusion of articles for a search period from start of each initial database to September 6, 2018.

anatomic location, output classes, AI models that were used, pretrained convolutional neural network (CNN), if applicable, size of the training set, size of the validation set or method, size of the test set, and performance measures (accuracy and AUC curve).

Output classes included fracture detection and/or classification. We considered fracture detection as a binary classifier with two inherent output classes (the presence of any fracture versus absence of a fracture). From what we could discern, these studies evaluated any type of fracture: both displaced fractures, which are easy to detect, and nondisplaced fractures, which can be subtler. Additionally, fracture classification addressed multiple output classes. For example, one study addressed a four-group classification system to distinguish among types of proximal humerus fractures (that is, the Neer classification [22]) [5], whereas another study addressed a subtype of femur fractures (AO-Type 32 [21]: a nine-group classification method for diaphyseal femur fractures ranging from simple spiral fractures to complex, irregular, comminuted fractures) [3].

Six studies described the use of a single AI model for detecting and/or classifying the fracture [5, 8, 14, 17, 30, 33], and four compared the performance of more than one model [1-3, 23].

We analyzed studies describing pretrained CNNs (AI models that were developed using large, separate datasets such as the ImageNet Large Scale Visual Recognition Challenge [25]) [5, 8, 14, 17, 23, 30] that were subsequently transferred to new datasets and AI models trained from scratch and implemented for new and unseen data.

Generally, two validation techniques are used to evaluate an AI model after the training phase: a subset of the dataset is retained as a validation set (that is, the size of the validation set) or a validation method is applied. The goal of using a validation set or validation method—especially in situations with small datasets—is to increase model robustness (for example, developing strategies to cope with errors during performance of a specific task). For example, k-fold cross-validation is a validation method that is applied to an automated computer-generated resampling procedure, in which a dataset is divided into smaller sets of different combinations (multiple folds or partitions), which allows it to train throughout many iterations. Although not mutually exclusive, each fold is iteratively used as a test set and the rest is used for training. The size of the test set is a partition of the dataset used for the final evaluation and determines the performance measures of the AI model.

The accuracy and AUC were assessed to provide information on each AI model in the test dataset because these were the most commonly addressed items (eight studies addressed accuracy [1-3, 5, 8, 14, 23, 30] and five studies addressed the AUC [5, 8, 14, 17, 30]). In our study, accuracy applied to the proportion of correctly detected or classified fractures and nonfractures to the total number of patients (such as the proportion of correct predictions over all cases). The AUC corresponds to the probability that a binary classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one [7]. An AUC of 1.0 would indicate perfect prediction, whereas 0.5 would indicate a prediction is no better than chance.

Distribution of Fracture Detection and Classification, Anatomical Location, Used AI Models, and Input Features

Nine studies addressed AI models for detecting fractures [1, 2, 5, 8, 14, 17, 23, 30, 33], whereas one study addressed fracture classification [3]. Chung et al. [5] were the only authors to report on both a fracture detection and fracture classification task (Table 1).

Anatomic fractures were located in the wrist [5, 14, 17, 23], hip [8, 30], spine [1, 33], ankle [23], diaphyseal femur [3], hand [23], and proximal humerus [5].

A pretrained CNN was the most frequently used AI model [5, 8, 14, 17, 23, 30], followed by neural networks [1-3], k-nearest neighbors [2, 3], support vector machines [3, 33], K-means [1], and linear discriminant analysis [3]. All AI models were supervised, except for the K-means, which is an unsupervised AI model.

Input features used in the AI models were as follows: eight studies used radiography as an imaging modality [2, 3, 5, 8, 14, 17, 23, 30], whereas two studies used CT [1, 33]. When radiography was the radiologic imaging modality, AP [5, 8, 23, 30] and lateral [5, 14, 17, 23] projections were most commonly used, followed by posteroanterior [17], oblique (two different types) [23], and scaphoid (four specific scaphoid views: proximal, distal, ulnar, and radial) [23].

Quality Appraisal

Ten studies were included. Quality appraisal demonstrated that the study aim was clear in seven studies (70%), possibly resulting in outcome bias for the remaining three studies (Fig. 3). In seven studies (70%), the inclusion and exclusion criteria for input features (all eligible radiographs and CT scans were included in the dataset) were clearly described, whereas selection bias could not be excluded for the remaining two studies (30%). Seven studies (70%) clearly described how they determined the ground truth (the reference standards in AI), subjecting the remainder to poorly trained AI models. All studies reported a clear distribution of the dataset (training, validation, and testing phases), described how the performance of an AI model was determined

Table 1. Studies evaluating the use of AI models in bone fracture detection and/or classification

Author, year	Input features	Imaging direction	Size dataset	Anatomical location	Ground truth label assignment*	Output classes	AI models used	Pretrained CNN	Size training set†	Size validation set‡ / validation method	Size test set	Performance (accuracy/ AUC)
Fracture detection												
Al-helo et al. [1]	CT	NA	50	Spine	NA	2	K-means	NA	100	NA	NA	98/NA
Al-helo et al. [1]	CT	NA	50	Spine	NA	2	NN	NA	90	5-LOCV	10	93/NA
Basha et al. [2]	X-ray	NA	180	NA	NA	2	NN	NA	78	NA	22	88/NA
Basha et al. [2]	X-ray	NA	180	NA	NA	2	kNN	NA	78	10-FCV	22	86/NA
Chung et al. [5]	X-ray	AP	1891	Proximal humerus	Combined approach§	2	CNN	ResNet-152	90	10-FCV	10	96/1.0
Gale et al. [8]	X-ray	AP	53279	Hip	Radiology reports	2	CNN	DenseNet	85.4	8.3	6.3	97/0.99
Kim and MacKinnon [14]	X-ray	Lat	1389	Distal radius and ulna	Radiology registrar‖	2	CNN	Inception v3	80	10	10	89/0.95
Lindsey et al. [17]	X-ray	PA, lat	34990	Wrist	Orthopaedic surgeon	2	CNN	U-Net	80	10	10	NA/0.97
Olczak et al. [23]	X-ray	AP, lat, obl, scaph	256458	Wrist, hand, ankle	Radiology reports	2	CNN	VGG_16	70	20	10	83/NA
Olczak et al. [23]	X-ray	AP, lat, obl, scaph	256458	Wrist, hand, ankle	Radiology reports	2	CNN	VGG_19	70	20	10	NA
Olczak et al. [23]	X-ray	AP, lat, obl, scaph	256458	Wrist, hand, ankle	Radiology reports	2	CNN	Network-in-network	70	20	10	NA
Olczak et al. [23]	X-ray	AP, lat, obl, scaph	256458	Wrist, hand, ankle	Radiology reports	2	CNN	VGG CNN S network	70	20	10	NA
Olczak et al. [23]	X-ray	AP, lat, obl, scaph	256458	Wrist, hand, ankle	Radiology reports	2	CNN	BVLC Reference CaffeNet	70	20	10	NA
Urakawa et al. [30]	X-ray	AP	3346	Intertrochanteric hip	Single radiologist	2	CNN	VGG_16	80	10	10	96/0.98
Yao et al. [33]	CT	NA	40	Thoracic and lumbar spine	Radiology reports	2	SVM	NA	50	10-FCV	50	
Fracture classification												
Bayram and Çakiroğlu [3]	X-ray	NA	196	Diaphyseal femur	NA	9	LibSVM	NA	100	10-FCV		90/NA
Bayram and Çakiroğlu [3]	X-ray	NA	196	Diaphyseal femur	NA	9	kNN	NA	100	10-FCV		87/NA
Bayram and Çakiroğlu [3]	X-ray	NA	196	Diaphyseal femur	NA	9	NN	NA	100	10-FCV		85/NA
Bayram and Çakiroğlu [3]	X-ray	NA	196	Diaphyseal femur	NA	9	LDA	NA	100	10-FCV		83/NA

Table 1. continued

Author, year	Input features	Imaging direction	Size dataset	Anatomical location	Ground truth label assignment*	Output classes	AI models used	Pretrained CNN	Size training set†	Size validation set‡ / validation method	Size test set	Performance (accuracy/ AUC)
Chung et al. [5]	X-ray	AP	1376	Proximal humerus	Combined approach§	4	CNN	ResNet-152	90	10-FCV	10	77/0.94

*The reference standard in AI.
†percentage of the total amount of the dataset.
‡percentage of the total amount included radiographs/CT scans.
§radiology registrar with 3 years of experience.
‖two shoulder orthopaedic surgeons and one radiologist assigned the ground truth labels with additional use of the corresponding CT-scan if consensus could not be achieved.
NA = not applicable (refers to anything not reported in a particular study).
CNN = convolutional neural network; AUC = area under the receiver operating characteristic curve; K-means = K refers to number of clusters generated by AI model; NN = neural network; 5-LOCV = leave-out cross-validation (in total 10 training rounds were performed, each round leaving five cases for validation); kNN = k-nearest neighbor; FCV = fold cross-validation (computer-generated resampling procedure, in which a dataset is divided into smaller sets of different combinations [for example, 10 folds or partitions], which allows it to train throughout many iterations); lat = lateral; front = frontal; obl = oblique (two different types); scaph = four specific scaphoid views (proximal, distal, ulnar, and radial); pa = posteroanterior; 20° = 20° degrees tilted lateral wrist view; SVM = support vector machine; LibSVM = library for support vector machine; LDA = linear discriminant analysis.

(accuracy and AUC), and clearly explained the AI model that was used (see Appendix; Supplemental Digital Content 2, http://links.lww.com/CORR/A195).

Statistical Analysis

Given the heterogeneity of the studies, we reported the range for accuracy and AUC for fracture detection and classification tasks. The sizes of the training, validation, and test sets are reported as percentages of the total number of the dataset.

There was no funding received to perform this work.

Results

AI Model Performance

Among the five studies using AUC for fracture detection AI had near perfect prediction (range, 0.95-1.0) [5, 8, 14, 17, 30]. The accuracy of fracture detection reported in seven studies ranged from 83% to 98% [1, 2, 5, 8, 14, 23, 30].

Seven studies addressed fracture detection on radiographs [2, 5, 8, 14, 17, 23, 30], and two studies addressed fracture detection on CT [1, 33].

In studies addressing fracture classification on radiographs, Chung et al. [5] found an AUC of 0.94 and an accuracy of 77% for classifying proximal humerus fractures into four groups (according to the Neer classification [22]). Bayram and Çakiroğlu [3] applied four AI models for classification of diaphyseal femur fractures into nine groups (AO-type 32 [21]) and found an accuracy ranging from 83% to 90%.

AI Models Compared with Humans

Three studies compared the performance of AI models with the performance of humans [5, 23, 30]. Urakawa et al. [30] used an AI model (that is, a pretrained CNN: VGG_16) for detecting hip fractures on an AP radiograph, which had a better AUC than five orthopaedic surgeons did (pretrained CNN: 0.98 [95% CI, 0.97-1.0] versus the five orthopaedic surgeons: 0.97 [95% CI, 0.95-0.97]; $p < 0.001$). Additionally, the difference in accuracy also favored the AI model (pretrained CNN: 96% [95% CI, 93-98] versus the five orthopaedic surgeons: 92% [95% CI, 89-95]; $p < 0.001$).

In a study by Olczak et al. [23], the accuracy of the best-performing AI model (a pretrained CNN: VGG_16) in detecting wrist, hand, and ankle fractures on several radiographic projections was equivalent to that of two senior orthopaedic surgeons (pretrained CNN: 83 [95% CI,

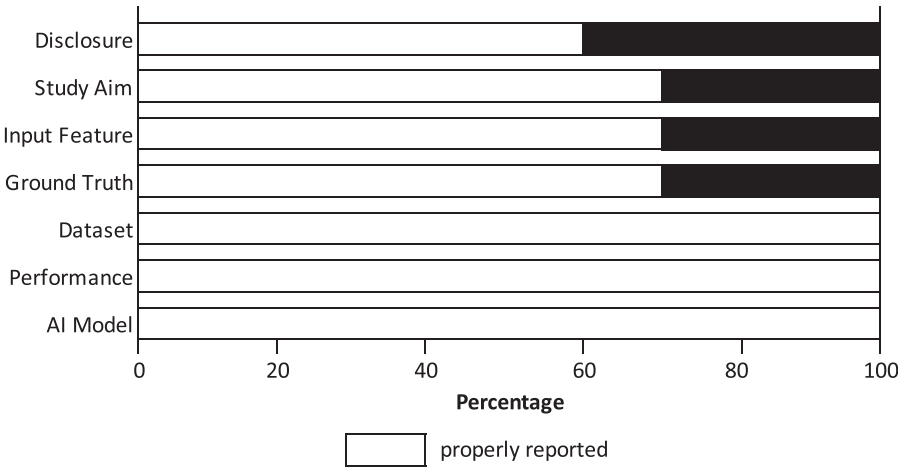

Fig. 3 We conducted a quality assessment of included studies using a seven-item checklist based on a modified methodologic index for nonrandomized studies (MINORS) instrument.

80-87] versus 82 [95% CI, 78-86] and 82 [95% CI, 78-85] for the two senior orthopaedic surgeons).

For detecting fracture, Chung et al. [5] used a pretrained CNN (Microsoft ResNet-152 ; Redmond, WA, USA) to detect proximal humerus fractures on an AP radiograph and compared the accuracy of the CNN with that of three human groups: general physicians (n = 28), general orthopaedists (n = 11), and an orthopaedist who specialized in the shoulder (n = 19). The accuracy of the AI model was superior to that of the human groups, although there was no statistical difference between the AI model and the general orthopaedist and shoulder orthopaedist groups (pretrained CNN: 96% [95% CI, 94-97] versus 85% [95% CI, 80-90] for the general physicians, 93% [95% CI, 90-96] for the general orthopaedists, and 93% [95% CI, 87-99] for the orthopaedists who specialized in the shoulder; $p < 0.001$). Additionally, except for one subset (greater tuberosity fractures), the pretrained CNN also demonstrated better accuracy for classifying proximal humerus fractures into four groups (according to the Neer classification [22]).

Discussion

AI can be used to develop predictive models based on large data sets. We analyzed the results of studies using AI for fracture detection and classification to determine the potential utility in fracture care. In a research setting, we found AI models are nearly as good as humans for detecting certain common fractures and—in two studies—outperformed humans for hip and proximal humerus fracture classification.

This study has several limitations. First, the studies addressed the performance of AI models based on only one projection when radiography was used as the input feature; this is in contrast to daily clinical practice, in which a surgeon bases his or her interpretation on multiple projections combined with taking the patient's history and performing a physical examination. AI models can be built to account for features of the interview, examination, and laboratory values (if applicable) along with image analysis. Second, the studies used a variety of approaches for assigning ground truth labels (the reference standard in AI) for each dataset with which the model was trained. For example, ground truth labels might be determined by a fellowship-trained musculoskeletal radiologist or through a thorough screening of reports in the medical record, consensus meeting among physicians with the additional use of more advanced imaging (such as CT images instead of radiographs) to resolve discrepancies, and radiologist reports. All these reference standards are subject to human error. AI models trained with more objective labeling assignments (for example, operative exposure) should result in more accurate and generalizable probabilities. Third, an appropriate risk of bias assessment tool does not exist for diagnostic studies. We therefore modified the methodologic index for nonrandomized studies (MINORS). Fourth, at present there are only a few preliminary studies used in simple diagnostic scenarios that may overestimate of the potential benefit of AI. Additional studies with clinically relevant settings will help evaluate the utility of AI. Fifth, although a broad search strategy encompassing three large databases was used, potentially relevant publications might have been missed. However, we deem this risk to be low, because we did not identify new eligible studies through screening the reference lists of included studies. In addition, we identified nine conference abstracts that have not been published yet, suggesting that AI is a developing research interest.

Our review found that AI was remarkably good at detecting common fractures. It is reasonable to assume that

the fracture locations were selected in these studies because they are common and yield large datasets. Most fractures in these areas are displaced and therefore relatively easy to detect by either a human or a computer. More subtle fractures (such as nondisplaced femoral neck or scaphoid fractures) need additional study as AI models might be less accurate. AI algorithms for diagnosing relatively obvious fractures might be useful for clinical scenarios where fractures might be overlooked (for example, multiple trauma) or in primary care or urgent care where a radiologist is not immediately available [11, 24], potentially replacing radiologists in this setting. AI could also be useful in difficult scenarios, such as suspected scaphoid or hip fractures, if proven to be accurate. A dispassionate examination of the probability of fracture could help surgeons and patients with decision-making. Further research should seek to identify situations in which AI could act in synergy with clinicians in fracture detection tasks, which are generally prone to misinterpretation or uncertainty. However, there are hurdles to overcome before implementation in clinical practice. First, a clinician might be reluctant to use a suggestion by an AI model since there is no human interface, it is not intuitive (complex statistical models), and it cannot be interrogated (the inscrutability of the magic "black box of AI"). The European Union has addressed liability concerns by incorporating a dictum in the General Data Protection Regulations that AI algorithmic decisions about humans must be interpretable and explainable [9]. Second, it remains debatable who would be held responsible if an algorithm errs and causes harm. Thus, appropriate legal regulations should be addressed before implementing AI into the clinical arena outside of research and quality improvement efforts. Lastly, most studies used datasets with ground truth labels that were based on formal reports from radiologists taken from the medical record to train the respective AI algorithms. For many reasons, these datasets have some inherent errors and misinterpretations. We may benefit from better ground truth labels (for example, operative findings or more sophisticated imaging) to develop more accurate AI algorithms.

AI had reasonable accuracy for classifying proximal humerus and diaphyseal femur fractures. Again, there is an issue with the lack of reference standards for the correct or most likely classification in these studies. For example, Chung et al. [5] determined the reference standard for the Neer Classification [22] by consensus of two shoulder surgeons and one radiologist using CT-images on occasion to reach agreement—an arguably inadequate reference standard for a classification that is known to be unreliable, even using CT scans [4, 19]. They also introduced selection bias by removing fractures for which consensus could not be reached. Alternatively, AI might use latent class analysis, a statistical technique that calculates the characteristics of diagnostic performance without a reference standard [16]. Bayesian inferences, another field of interest proposed by Kim and MacKinnon [14] could be used to produce more meaningful predictions that accurately reflects the probable outcome, by accounting for the influence of fracture incidence when analysing accuracy [14].

Two studies found that AI was better than humans at detecting and classifying hip and proximal humerus fractures, and one found equivalent performance for detecting wrist, hand, and ankle fractures [5, 23, 30]. This suggests that—at least for relatively straightforward diagnostic scenarios—AI can be useful. There are important gaps to consider. These studies based their ground truth on human assessment (for example, radiology reports or a single radiologist's interpretation) [5, 23, 30]. As clinicians are susceptible to error, the AI models were trained and tested with images that had some level of inaccuracy [12]. As such, AI models might erroneously report good performance, while this would not be detected as a diagnostic error by the model. Additionally, these AI models can diagnose the fracture, but cannot discern which fractures may involve a bone tumor, for example. In contrast, an orthopaedic surgeon or radiologist is more likely to detect additional relevant findings when evaluating radiographs of fractures. Moreover, physicians are able to combine patients' preferences and objective parameters (such as laboratory values) into careful clinical decision-making.

The current thinking about AI application in medicine seems to be that narrow tasks with predefined context are most suitable, such as recognizing the border of an organ to suggest where to stop scanning or detecting suspicious areas in an image [28]. Risk prediction and therapeutics are more challenging for AI. A lack of reliable and accurate standards on which to train and test an algorithm for certain disease entities (such as delirium), makes the probabilities generated by AI less suitable and applicable for the end-user. Furthermore, an algorithm's output is only an association, not a causative relationship [18]. Therefore, physicians should always balance the probable outcome of this output and decide whether it applies to a specific patient. According to Verghese et al. [31], AI applications and clinicians should always cooperate: AI helps predict and the clinician compassionately explains and decides.

We speculate that AI might outperform humans for many probabilistic tasks that are based on data. However, the largest challenges will be to find ways to collect and analyze large amounts of data efficiently and to overcome legal issues. Despite the current shortcomings, such as inadequate ground truth label assignment, we believe that physicians will benefit by embracing AI rather than ignoring or dismissing it. For fracture care, these models might aid surgeons by drawing their attention to fractures or fracture characteristics that could cause harm if overlooked. Future studies in this area might focus on AI as a

tool to assist with complex and uncertain clinical tasks (for example, determining the response of bone tumors to chemotherapy, or detecting nondisplaced or occult fractures) and in decision support.

Acknowledgments None.

References

1. Al-Helo S, Alomari RS, Ghosh S, Chaudhary V, Dhillon G, Al-Zoubi MB, Hiary H, Hamtini TM. Compression fracture diagnosis in lumbar: a clinical CAD system. *Int J Comput Assist Radiol Surg*. 2013;8:461-469.
2. Basha CMAKZ, Padmaja M, Balaji GN. Computer Aided Fracture Detection System. *J Med Imaging Health Inform*. 2018; 8:526-531.
3. Bayram F, Çakiroğlu M. DIFFRACT: DIaphyseal Femur FRActure Classifier SysTem. *Biocybern Biomed Eng*. 2016;36: 157-171.
4. Carofino BC, Leopold SS. Classifications in brief: the Neer classification for proximal humerus fractures. *Clin Orthop Relat Res*. 2013;471:39-43.
5. Chung SW, Han SS, Lee JW, Oh KS, Kim NR, Yoon JP, Kim JY, Moon SH, Kwon J, Lee HJ, Noh YM, Kim Y. Automated detection and classification of the proximal humerus fracture by using deep learning algorithm. *Acta Orthop*. 2018;89:468-473.
6. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542:115-118.
7. Fawcett T. An introduction to ROC analysis. *Pattern Recognition Letters*. 2006;27:861-874.
8. Gale W, Oakden-Rayner L, Carneiro G, Bradley AP, Palmer LJ. Detecting hip gractures with tadiologist-level performance using deep neural networks. Available at: https://arxiv.org/abs/1711.06504. Accessed November 17, 2017.
9. Goodman B, Flaxman S. European Union regulations on algorithmic decision-making and a "right to explanation". Available at: https://arxiv.org/abs/1606.08813. Accessed June 28, 2016.
10. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, Venugopalan S, Widner K, Madams T, Cuadros J, Kim R, Raman R, Nelson PC, Mega JL, Webster DR. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316:2402-2410.
11. Guly HR. Diagnostic errors in an accident and emergency department. *Emerg Med J*. 2001;18:263-269.
12. Hallas P, Ellingsen T. Errors in fracture diagnoses in the emergency department–characteristics of patients and diurnal variation. *BMC Emerg Med*. 2006;6:4.
13. Karhade AV, Thio Q, Ogink PT, Shah AA, Bono CM, Oh KS, Saylor PJ, Schoenfeld AJ, Shin JH, Harris MB, Schwab JH. Development of machine learning algorithms for prediction of 30-day mortality after surgery for spinal metastasis. *Neurosurgery*. 2018.
14. Kim DH, MacKinnon T. Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks. *Clin Radiol*. 2018;73:439-445.
15. Kooi T, Litjens G, van Ginneken B, Gubern-Merida A, Sanchez CI, Mann R, den Heeten A, Karssemeijer N. Large scale deep learning for computer aided detection of mammographic lesions. *Med Image Anal*. 2017;35:303-312.
16. LaJoie AS, McCabe SJ, Thomas B, Edgell SE. Determining the sensitivity and specificity of common diagnostic tests for carpal tunnel syndrome using latent class analysis. *Plast Reconstr Surg*. 2005;116:502-507.
17. Lindsey R, Daluiski A, Chopra S, Lachapelle A, Mozer M, Sicular S, Hanel D, Gardner M, Gupta A, Hotchkiss R, Potter H. Deep neural network improves fracture detection by clinicians. *Proc Natl Acad Sci U S A*. 2018;45:11591-11596.
18. Maddox TM, Rumsfeld JS, Payne PRO. Questions for Artificial Intelligence in Health Care. *JAMA*. 2019;321:31-32.
19. Majed A, Macleod I, Bull AM, Zyto K, Resch H, Hertel R, Reilly P, Emery RJ. Proximal humeral fracture classification systems revisited. *J Shoulder Elbow Surg*. 2011;20:1125-1132.
20. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med*. 2009;6:e1000097.
21. Müller M. *The Comprehensive Classification of Fractures of Long Bones*. New York, NY: Springer-Verlag. 1990.
22. Neer CS, 2nd. Displaced proximal humeral fractures. I. Classification and evaluation. *J Bone Joint Surg Am*. 1970;52: 1077-1089.
23. Olczak J, Fahlberg N, Maki A, Razavian AS, Jilert A, Stark A, Skoldenberg O, Gordon M. Artificial intelligence for analyzing orthopedic trauma radiographs. *Acta Orthop*. 2017;88: 581-586.
24. Pfeifer R, Pape HC. Missed injuries in trauma patients: A literature review. *Patient Saf Surg*. 2008;2:20.
25. Russakovsky O, Olga R, Jia D, Hao S, Jonathan K, Sanjeev S, 115. ImageNet large scale visual recognition challenge. *Int J Comput Vis*. 2015:211–252.
26. Samuel AL. Some studies in machine learning using the game of checkers. *IBM J Res Dev*. 1959;3:210-220.
27. Slim K, Nini E, Forestier D, Kwiatkowski F, Panis Y, Chipponi J. Methodological index for non-randomized studies (minors): development and validation of a new instrument. *ANZ J Surg*. 2003;73:712-716.
28. Stead WW. Clinical Implications and Challenges of Artificial Intelligence and Deep Learning. *JAMA*. 2018;320:1107-1108.
29. Thio Q, Karhade AV, Ogink PT, Raskin KA, De Amorim Bernstein K, Lozano Calderon SA, Schwab JH. Can machine-learning techniques be used for 5-year survival prediction of patients with chondrosarcoma? *Clin Orthop Relat Res*. 2018;476: 2040-2048.
30. 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*. 2018;48:239-244.
31. Verghese A, Shah NH, Harrington RA. What This Computer Needs Is a Physician: Humanism and Artificial Intelligence. *JAMA*. 2018;319:19-20.
32. Wells GA, Shea B, O'Connell D, Peterson J, Welch V, Losos M, Tugwell P. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. Avalaible at: http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp.
33. Yao J, Burns JE, Munoz H, Summers RM. Cortical shell unwrapping for vertebral body abnormality detection on computed tomography. *Comput Med Imaging Graph*. 2014;38: 628-638.