Clin Orthop Relat Res (2020) 478:2653-2659
DOI 10.1097/CORR.0000000000001318

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

Clinical Research

Is Deep Learning On Par with Human Observers for Detection of Radiographically Visible and Occult Fractures of the Scaphoid?

David W. G. Langerhuizen MD, Anne Eva J. Bulstra MD, Stein J. Janssen 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: 18 October 2019 / Accepted: 30 April 2020 / Published online: 19 May 2020

Abstract

Background Preliminary experience suggests that deep learning algorithms are nearly as good as humans in detecting common, displaced, and relatively obvious fractures (such as, distal radius or hip fractures). However, it is not known whether this also is true for subtle or relatively nondisplaced fractures that are often difficult to see on radiographs, such as scaphoid fractures.

Questions/purposes (1) What is the diagnostic accuracy, sensitivity, and specificity of a deep learning algorithm in detecting radiographically visible and occult scaphoid fractures using four radiographic imaging views? (2) Does adding patient demographic (age and sex) information improve the diagnostic performance of the deep learning algorithm? (3) Are orthopaedic surgeons better at diagnostic accuracy, sensitivity, and specificity compared with deep learning? (4) What is the interobserver reliability among five human observers and between human consensus and deep learning algorithm?

One of the authors (DR) certifies that he received payments in the amount of USD 10,000 to USD 100,000 in royalties from Skeletal Dynamics (Miami, FL, USA) and payments in the amount of less than USD 10,000 in royalties 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.
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.
Each author certifies that his or her institution approved or waived approval for the human protocol for this investigation and that all investigations were conducted in conformity with ethical principles of research.
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.
This study was performed at Flinders Medical Centre, Adelaide, Australia and the Amsterdam University Medical Centre, Amsterdam, The Netherlands.

D. W. G. Langerhuizen, S. J. Janssen, G. M. M. J. Kerkhoffs, Department of Orthopaedic Surgery, Amsterdam Movement Sciences (AMS), Amsterdam University Medical Centre, Amsterdam, The Netherlands

A. E. J. Bulstra, R. L. Jaarsma, J. N. Doornberg, Flinders University, Department of Orthopaedic & Trauma Surgery, Flinders Medical Centre, Adelaide, Australia

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 Surgery, Amsterdam University Medical Centre, AMC location, University of Amsterdam Meibergdreef 9, 1105AZ, Amsterdam, the Netherlands, Email: david.langerhuizen@gmail.com

Wolters Kluwer

Methods We retrospectively searched the picture archiving and communication system (PACS) to identify 300 patients with a radiographic scaphoid series, until we had 150 fractures (127 visible on radiographs and 23 only visible on MRI) and 150 non-fractures with a corresponding CT or MRI as the reference standard for fracture diagnosis. At our institution, MRIs are usually ordered for patients with scaphoid tenderness and normal radiographs, and a CT with radiographically visible scaphoid fracture. We used a deep learning algorithm (a convolutional neural network [CNN]) for automated fracture detection on radiographs. Deep learning, an advanced subset of artificial intelligence, combines artificial neuronal layers to resemble a neuron cell. CNNs—essentially deep learning algorithms resembling interconnected neurons in the human brain—are most commonly used for image analysis. Area under the receiver operating characteristic curve (AUC) was used to evaluate the algorithm's diagnostic performance. An AUC of 1.0 would indicate perfect prediction, whereas 0.5 would indicate that a prediction is no better than a flip of a coin. The probability of a scaphoid fracture generated by the CNN, sex, and age were included in a multivariable logistic regression to determine whether this would improve the algorithm's diagnostic performance. Diagnostic performance characteristics (accuracy, sensitivity, and specificity) and reliability (kappa statistic) were calculated for the CNN and for the five orthopaedic surgeon observers in our study.

Results The algorithm had an AUC of 0.77 (95% CI 0.66 to 0.85), 72% accuracy (95% CI 60% to 84%), 84% sensitivity (95% CI 0.74 to 0.94), and 60% specificity (95% CI 0.46 to 0.74). Adding age and sex did not improve diagnostic performance (AUC 0.81 [95% CI 0.73 to 0.89]). Orthopaedic surgeons had better specificity (0.93 [95% CI 0.93 to 0.99]; $p < 0.01$), while accuracy (84% [95% CI 81% to 88%]) and sensitivity (0.76 [95% CI 0.70 to 0.82]; $p = 0.29$) did not differ between the algorithm and human observers. Although the CNN was less specific in diagnosing relatively obvious fractures, it detected five of six occult scaphoid fractures that were missed by all human observers. The interobserver reliability among the five surgeons was substantial (Fleiss' kappa = 0.74 [95% CI 0.66 to 0.83]), but the reliability between the algorithm and human observers was only fair (Cohen's kappa = 0.34 [95% CI 0.17 to 0.50]).

Conclusions Initial experience with our deep learning algorithm suggests that it has trouble identifying scaphoid fractures that are obvious to human observers. Thirteen false positive suggestions were made by the CNN, which were correctly detected by the five surgeons. Research with larger datasets—preferably also including information from physical examination—or further algorithm refinement is merited.

Level of Evidence Level III, diagnostic study.

Introduction

Deep learning gained great appeal when Google's DeepMind computer defeated the world's number one Go player [1]. Deep learning, an advanced subset of artificial intelligence, combines artificial neuronal layers to resemble a neuron cell. Essentially, these algorithms—highly complex mathematical models—derive rules and patterns from data to estimate the probability of a diagnosis or outcome without human intervention. These algorithms can be applied to imaging tasks such as skin cancer detection on photographs or detection of critical findings in head CT scans [2, 5].

Using different data set sizes, initial experience with fracture detection on radiographs suggests that deep learning algorithms are (nearly) as good as humans at detecting certain common fractures such as distal radius, proximal humerus, and hip fractures [11]. However, many of those fractures are displaced and relatively obvious on radiographs.

It is known that scaphoid fractures can have long-term consequences if not properly diagnosed. A previous study applied five deep learning algorithms to detect wrist, hand (including scaphoid), and ankle fractures; however, they did not report their algorithm's performance for scaphoid fractures specifically [13]. As such, it is not yet clear whether deep learning algorithms will be useful for the detection of relatively subtle and often radiographically invisible nondisplaced femoral neck or scaphoid fractures that are often overlooked by humans, particularly nonspecialists [9].

Therefore, we asked: (1) What is the diagnostic accuracy, sensitivity, and specificity of a deep learning algorithm in detecting radiographically visible and occult scaphoid fractures using four radiographic imaging views? (2) Does adding patient demographic (age and sex) information improve the diagnostic performance of the deep learning algorithm? (3) Are orthopaedic surgeons better at diagnostic accuracy, sensitivity, and specificity compared with deep learning? (4) What is the interobserver reliability among five human observers and between human consensus and deep learning algorithm?

Patients and Methods

Data Set and Pre-processing

Our institutional review board approved this retrospective study. Our institution still uses a paper medical record, which makes it difficult to search for patients with specific diagnoses and tests. The picture archiving and communication system (PACS) is electronic and easier to search. We

used two strategies to identify at least 300 scaphoid series of radiographs.

The first strategy was based on the fact that clinicians in our institution usually order an MRI in patients with suspected scaphoid fractures and normal radiographs and a CT with radiographically visible scaphoid fracture. This strategy identified MRI and CT of the scaphoid and then sought corresponding radiographs of scaphoid fractures. We searched the PACS database using the terms "MR scaph", "CT hand", "CT wrist", and "CT extr" and identified 326 patients: 150 that were excluded because the radiographs were incomplete or distorted by cast or splint materials and 176 with adequate radiographic scaphoid series including 13 MRI-confirmed fractures, 59 CT-confirmed fractures, and 104 MRI-confirmed nonfractures.

In the second strategy, we searched PACS for "Xr scaph" and searched them one by one for a corresponding MRI or CT image and an adequate series of radiographs not distorted by plaster. We found 124 additional patients including 10 with MRI-confirmed fractures, 68 with CT-confirmed fractures, 46 MRI-confirmed nonfractures, and 17 CT-confirmed nonfractures. Two observers (DWGL, AEJB) used this strategy to identify patients until we had 150 radiographs of scaphoids with a fracture (127 visible on radiographs and 23 only visible on MRI) and 150 without a fracture, numbers chosen before starting the search and based on typical training strategies. Age and sex demographics were provided by PACS. The mean age at diagnosis was 36 years (SD 16), and 62% (185 of 300) of patients were male. We randomly divided the dataset into a train, a validation, and a test group (180:20:100), each divided 50:50 by presence of a fracture. The radiographically invisible fractures were randomly and evenly distributed between the three groups. To match the predefined image size of the deep learning framework (Fig. 1), we manually cropped and resized all Digital Imaging and Communications in Medicine (DICOM) files into a 350 x 300 pixels rectangle capturing the scaphoid (see Appendix 1; Supplemental Digital Content 1, http://links.lww.com/CORR/A353). By automatically rotating, zooming, changing height/width, and horizontal/vertical flipping, all preformatted images were 10-fold augmented with the intent to increase robustness of the algorithm.

Algorithm: Convolutional Neural Network

Convolutional neural networks (CNNs) are complex algorithms resembling interconnected neurons in the human brain. CNNs are a form of deep learning commonly used to analyze images. In deep learning, the computer analyzes both features that are recognizable to humans (for example, the eyes or the nose) and features that are not recognizable to humans (such as edges or transitions). A CNN learns by developing and testing algorithms again and again (in iterations) until it has optimized its ability to identify the feature assigned: in this case, fracture of the scaphoid. When approaching a new image recognition task, it can be helpful to start with a CNN that is already trained to identify features in images. We used an open-source pretrained CNN (Visual Geometry Group, Oxford, United Kingdom [18]) trained on more than 1 million non-medical images with 1000 object categories [16] (see Appendix 2; Supplemental Digital Content 2, http://links.lww.com/CORR/A354).

A test group of 100 images was randomly selected for use in the tests to determine the algorithm performance. We evaluated the model using the following performance metrics: area under the receiving operating characteristic (AUC) curve, accuracy, sensitivity, and specificity. We set the diagnostic cutoff point at a value that maximized sensitivity, at the cost of a slightly decreased specificity [3, 8, 9].

Codes were written in Python Version 3.6.8 (Python Software Foundation, Wilmington, DE, USA) with the packages scikit-learn (0.20.3) and TensorFlow (1.13.1).

Human Observers

We compared the performance metrics of the model with five surgeons (RLJ, JND, MMAJ, NK, JWW). Three orthopaedic trauma surgeons (16, 3, and 2 years after completion of residency training) and two upper limb surgeons (25 and 2 years after completion residency training) each reviewed the same 100 patients as the model. In our hospital, upper limb surgeons deliver care for the entire upper extremity. The surgeons were not aware of the total number of fracture and nonfracture patients in the test set. All fractures were presented as uncropped Digital Imaging and Communications in Medicine (DICOM) files, which we loaded into Horos (version 3.3.4, Annapolis, MD, USA). Surgeons were asked to identify the presence or absence of a scaphoid fracture on four radiographic views. Again, we calculated the accuracy, sensitivity, and specificity for each surgeon as well as the mean among surgeons for each measure to compare with the CNN.

Statistical Analysis

Continuous variables were presented with mean and SD and categorical variables with frequencies and percentages.

Accuracy is defined as the proportion of correctly detected cases among all cases. The AUC reflects the probability that a binary classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one [6]. An AUC of 1.0 corresponds to perfect

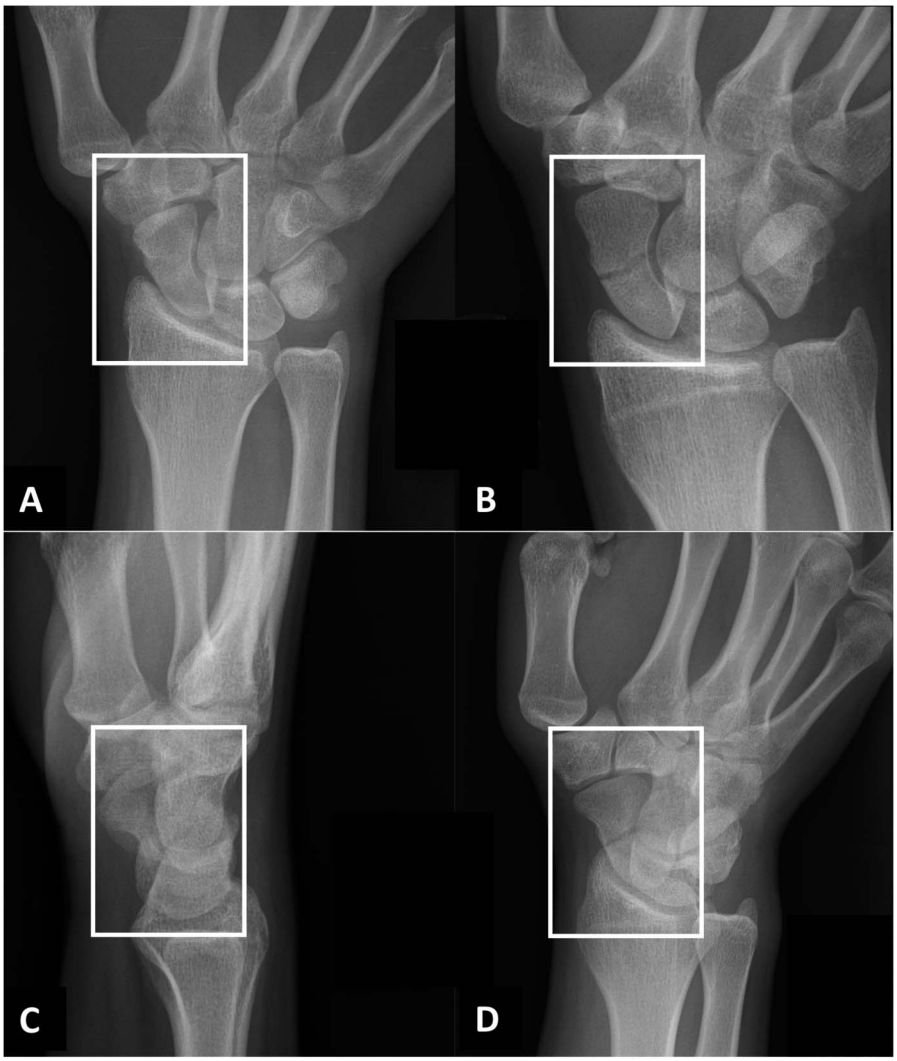

Fig. 1 A-D A radiographic scaphoid fracture series for patients with a clinical suspicion for scaphoid fracture at our hospital. The following four projections were fed into the deep learning framework: (**A**) posterior-anterior ulnar deviation; (**B**) uptilt (that is, an elongated view with tube angle adjusted over 30°); (**C**) lateral; and (**D**) 45° oblique projections. The white boxes illustrate the cropped and resized radiographs (350 x 300 pixels) that are fed into the deep learning framework (VGG 16).

classification, whereas 0.5 indicates a prediction equal to chance. Sensitivity corresponds to the proportion of correctly identified fractures among all actual fractures, while specificity refers to the proportion of correctly identified nonfractures among all nonfractures. We calculated 95% confidence intervals using a Z-score of 1.96. Overlapping 95% CIs indicate no significant difference. A McNemar's test was used to compare sensitivity and specificity between the algorithm and human observers. The probability of a scaphoid fracture generated by the CNN, sex, and age were included in a multivariable logistic regression to determine whether this would improve the algorithm's diagnostic performance.

Kappa, which is a chance-corrected measure, corresponds to the agreement among observers. We used Fleiss' kappa to determine interobserver reliability among surgeons for evaluating the presence or absence of scaphoid fractures. We used Cohen's kappa to calculate reliability between the CNN and majority vote of human observers. According to Landis and Koch [10], a kappa between 0.21 and 0.40 reflects fair agreement, a kappa between 0.41 and 0.60 indicates moderate agreement, a kappa between 0.61 and 0.80 reflects substantial agreement, while a kappa above 0.80 indicates almost perfect agreement .

We performed statistical analyses using Stata 15.0 (StataCorp LP, College Station, TX, USA) and RStudio (Boston, MA, USA) with the packages CalibrationCurves, ggplot2, grid, and precrec.

There were no missing data.

Results

Performance of CNN

For detection of scaphoid fractures among suspected scaphoid fractures, the CNN reported an AUC of 0.77 (95% CI 0.66 to 0.85) (Fig. 2). The CNN correctly detected 72 of 100 patients (accuracy 72% [95% CI 60% to 84%]). Eight of 50 confirmed scaphoid fractures were not identified (sensitivity 0.84 [95% CI 0.74 to 0.94]), while 20 of 50 patients without a fracture were incorrectly diagnosed as having a fracture of the scaphoid (specificity 0.60 [95% CI 0.46 to 0.74]).

Performance of CNN Combined with Patient Demographics

Combining age and sex with the generated probabilities of the CNN did not improve the AUC (0.81; 95% CI 0.73 to 0.89). The output of this model was converted into a formula for calculating the probability of a fracture (see Appendix 3; Supplemental Digital Content 3, http://links.lww.com/CORR/A355).

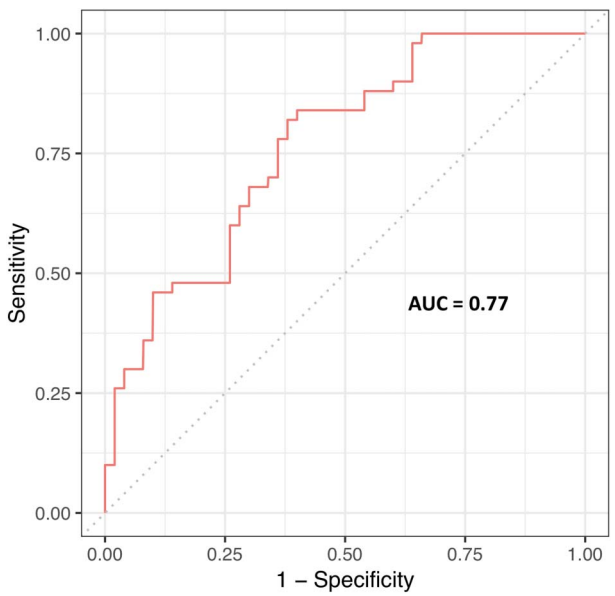

Fig. 2 This figure depicts the receiver operating curve for the CNN at the optimal diagnostic cutoff point (0.37).

Performance of CNN Compared with Human Observers

Specificity favored the human observers (five orthopaedic surgeons 0.93 [95% CI 0.87 to 0.99] versus CNN 0.60 [95% CI 0.46 to 0.74]; $p < 0.01$). Accuracy for distinguishing between scaphoid fractures and nonfractures was comparable between human observers and the CNN (five orthopaedic surgeons 84% [95% CI 81% to 88%] versus CNN 72% [95% CI 60 to 84]) (Table 1). Sensitivity was also comparable between the CNN and human observers (five orthopaedic surgeons: 0.76 [95% CI 0.70 to 0.82]) versus CNN: 0.84 [95% CI 0.74 to 0.94]; $p = 0.29$).

Six scaphoid fractures were missed by all surgeons and therefore considered occult fractures. The CNN detected five of six occult scaphoid fractures. In addition, five human observers detected three fractures that were missed by the CNN. Two fractures, diagnosed by four of five human observers, were also missed by the CNN. In contrast, thirteen false positive suggestion of the CNN, were correctly detected by the surgeons.

The Interobserver Reliability of Human Observers

Interobserver agreement between five surgeons was higher than between human consensus and the algorithm (0.74 [95% CI 0.66 to 0.83] versus 0.34 [95% CI 0.17 to 0.50]) (Table 2).

Discussion

In medicine, deep learning has primarily been applied to image analysis. In a research setting, use of deep transfer learning showed promising performance for fracture detection and classification for relatively straightforward clinical scenarios [11]. It is not yet clear that deep learning will be useful for radiographic fracture detection in scenarios where fractures are often overlooked by human observers. Using a relatively small data set of 300 patients, our deep learning algorithm demonstrated a moderate better overall performance for detection of radiographically visible and occult fractures (AUC 0.77 [95% CI 0.66 to 0.85]) and human observers had notably better specificity. The algorithm might have performed better if provided with more data.

This study has several limitations. First, we selected our patients from readily available and searchable radiology reports and intentionally introduced a spectrum bias by collecting 150 MRI- or CT-confirmed fractures and 150 confirmed nonfractures. Although this was needed to sufficiently train the algorithm, readers should keep in mind that our data set does not represent the true prevalence of

Table 1. A comparison of performance metrics between the CNN and the mean of five orthopaedic surgeons

Diagnostic performance characteristic	Orthopaedic surgeons	CNN[a]	p value
Accuracy (95% confidence interval)	84% (81% to 88%)	72% (60% to 84%)	[b]
Sensitivity (95% CI)	0.76 (0.70 to 0.82)	0.84 (0.74 to 0.94)	0.29
Specificity (95% CI)	0.93 (0.87 to 0.99)	0.60 (0.46 to 0.74)	**< 0.01**

[a]CNN = convolutional neural network at cutoff point 0.37.
[b]We did not calculate a p value, since McNemar's test is sensitive to the proportion of fractures as well as nonfractures.
Bold indicates statistical significance ($p < 0.05$).

radiographic scaphoid fracture appearance. Second, we were only able to include 300 patients because we could only search a 9-year period starting in January 2010. Three hundred radiographs is a relatively small sample size for deep learning, but more than adequate for logistic regression. A larger data set might improve the diagnostic performance of the CNN. We cannot be certain because, to this point, there is no consensus on a priori sample size in deep learning. It depends on the specific image-analysis task, the quality of the data set, the programming techniques used, and type of deep learning algorithm applied [14]. Third, the ground truth labels (that is, the reference standard diagnosis of scaphoid fracture or not) are based on radiologist interpretations of CT or MRI images, which have limited reliability and untestable accuracy. Given the small number of MRIs with diagnosed fracture and CT with diagnosed nonfractures, we believe any misdiagnoses would have little influence on the model. Fourth, radiographs were manually cropped and resized by one investigator (DWGL), which might introduce bias. However, given that cropping was assisted by an easy-to-use program scripted in Python, we feel it is very likely that another investigator would resize the images similarly. But, one should keep in mind that cropped radiographs may not reflect a clinical scenario, as other potentially relevant findings in a real-size radiograph were not assessable (such as, concomitant fractures or scapholunate dissociation). Furthermore, irrelevant regions in a radiograph were removed and therefore not evaluated by the model. A more in-depth deep learning framework, accounting for the entire wrist radiograph, merits further study. For now, the memory capacity of graphics processing units limits the usable image size. Fifth, among the five human observers, two surgeon raters treated some of the patients in the study, which might have influenced their diagnoses. We feel this would have negligible influence on our findings. Sixth, although incorporating injury details, signs, and symptoms would have been of interest to incorporate in a logistic regression model as it typical for a clinical prediction rule, they were not commonly reported in a patient's medical record. CNNs only evaluate images, but the probabilities generated can be included in clinical prediction rules.

The AUC of the CNN for detection of scaphoid fractures is not good enough to replace human observers or more sophisticated imaging, but it does suggest the potential to be used as a pre-screen or clinical prediction rule for triage of suspected scaphoid fractures that might benefit additional imaging. Displaced proximal humerus, distal radius, and intertrochanteric hip fractures are relatively easy to detect and not a good test of the potential utility of artificial intelligence [3, 9, 17]. Subtle and invisible fractures may be more of a challenge. Prior studies using deep learning algorithms to detect radiographically subtle hip and distal radius fractures had better performance than our model [7, 9, 12, 17]. Larger data sets, use of other pretrained CNNs, varying degrees of algorithm refinement and hyper-parameter tuning, as well as other anatomical fracture locations may explain why these studies differ with our findings. Also, we might not have had sufficient images to train the upper layers of the pretrained CNN.

Adding sex and age did not improve diagnostic performance. Future research might investigate whether

Table 2. Contingency table comparing prediction of convolutional neural network to human consensus (agreement ≥ three surgeons)

		Fracture (n = 50)	Non-fracture (n = 50)
Fracture (predicted)	CNN	42	20
	Human consensus	38	1
Non-fracture (predicted)	CNN	8	30
	Human consensus	12	49

CNN = convolutional neural network

incorporating computer analysis of images improves performance of clinical prediction rules that include demographics, injury details, symptoms, and signs to better triage the use of MRI as well as increase its diagnostic performance by increasing the pretest odds of a fracture [4, 15]. The pretest odds could be increased with CNNs, clinical prediction rules, or a combination of both.

Our deep learning algorithm was less specific than human observers but detected five of six occult fractures in the test dataset. On the other hand, caution is warranted because the CNN missed some radiographically visible fractures.

The finding that reliability of fracture diagnosis was substantial (0.74) for the five orthopaedic surgeons and only fair (0.34) between the surgeons and the CNN we interpret as a reflection of the difficulty the deep learning algorithm has with detecting radiographically visible fractures. At the diagnostic cutoff point—chosen to maximize sensitivity—the algorithm's specificity was considerably lower compared with human observers. A different cutoff point may have resulted in more or less the same reliability for detecting scaphoid fractures. It may go without saying that CNNs are known for being highly complex and, to date, not intuitive for the end-user. It is therefore not possible to understand how a CNN reaches its suggestion.

In conclusion, using a relatively small dataset, a deep learning algorithm was inferior to human observers at identifying scaphoid fractures on radiographs. Further study may help evaluate whether a larger dataset and algorithm refinement can increase the performance of deep learning for the diagnosis of scaphoid fractures, some of which are radiographically invisible. In addition, incorporating predictions from a deep-learning algorithm into clinical prediction rules that also account for demographics, injury details, symptoms, and signs merits further study.

Acknowledgments We thank the following orthopaedic surgeons for their participation: M. M. A. Janssen MD, PhD, N. Kruger MD, and J. W. White MBBS, PhD.

References

1. British Broadcasting Corporation. Artificial intelligence: Go master Lee Se-dol wins against AlphaGo program. Available at: https://www.bbc.com/news/technology-35797102. Accessed March 13, 2016.
2. Chilamkurthy S, Ghosh R, Tanamala S, Biviji M, Campeau NG, Venugopal VK, Mahajan V, Rao P, Warier P. Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. *Lancet (London, England)*. 2018;392: 2388-2396.
3. 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.
4. Duckworth AD, Buijze GA, Moran M, Gray A, Court-Brown CM, Ring D, McQueen MM. Predictors of fracture following suspected injury to the scaphoid. *J Bone Joint Surg Br*. 2012;94: 961-968.
5. 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.
6. Fawcett T. An introduction to ROC analysis. *Pattern Recognition Letters*. 2006;27:861-874.
7. Gale W, Oakden-Rayner L, Carneiro G, Bradley AP, Palmer LJ. Detecting Hip Fractures with Radiologist-Level Performance Using Deep Neural Networks. 2017; Available at: https://arxiv.org/abs/1711.06504. Accessed November 17, 2017.
8. Gan K, Xu D, Lin Y, Shen Y, Zhang T, Hu K, Zhou K, Bi M, Pan L, Wu W, Liu Y. Artificial intelligence detection of distal radius fractures: a comparison between the convolutional neural network and professional assessments. *Acta Orthop*. 2019;90: 394-400.
9. Kim DH, MacKinnon T. Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks. *Clin Radiol*. 2018;73:439-445.
10. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33:159-174.
11. Langerhuizen DWG, Janssen SJ, Mallee WH, van den Bekerom MPJ, Ring D, Kerkhoffs G, Jaarsma RL, Doornberg JN. 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.
12. 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.
13. 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.
14. Ranschaert ER. *Artificial Intelligence in Medical Imaging*. eBook. Switzerland, AG: Springer. 2019.
15. Rhemrev SJ, Beeres FJ, van Leerdam RH, Hogervorst M, Ring D. Clinical prediction rule for suspected scaphoid fractures: A prospective cohort study. *Injury*. 2010;41:1026-1030.
16. 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.
17. 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.
18. Zhong S, Li K, Feng R. Deep Convolutional Hamming Ranking Network for Large Scale Image Retrieval. Available at: https://ieeexplore.ieee.org/document/7052856. Accessed August 19, 2016.